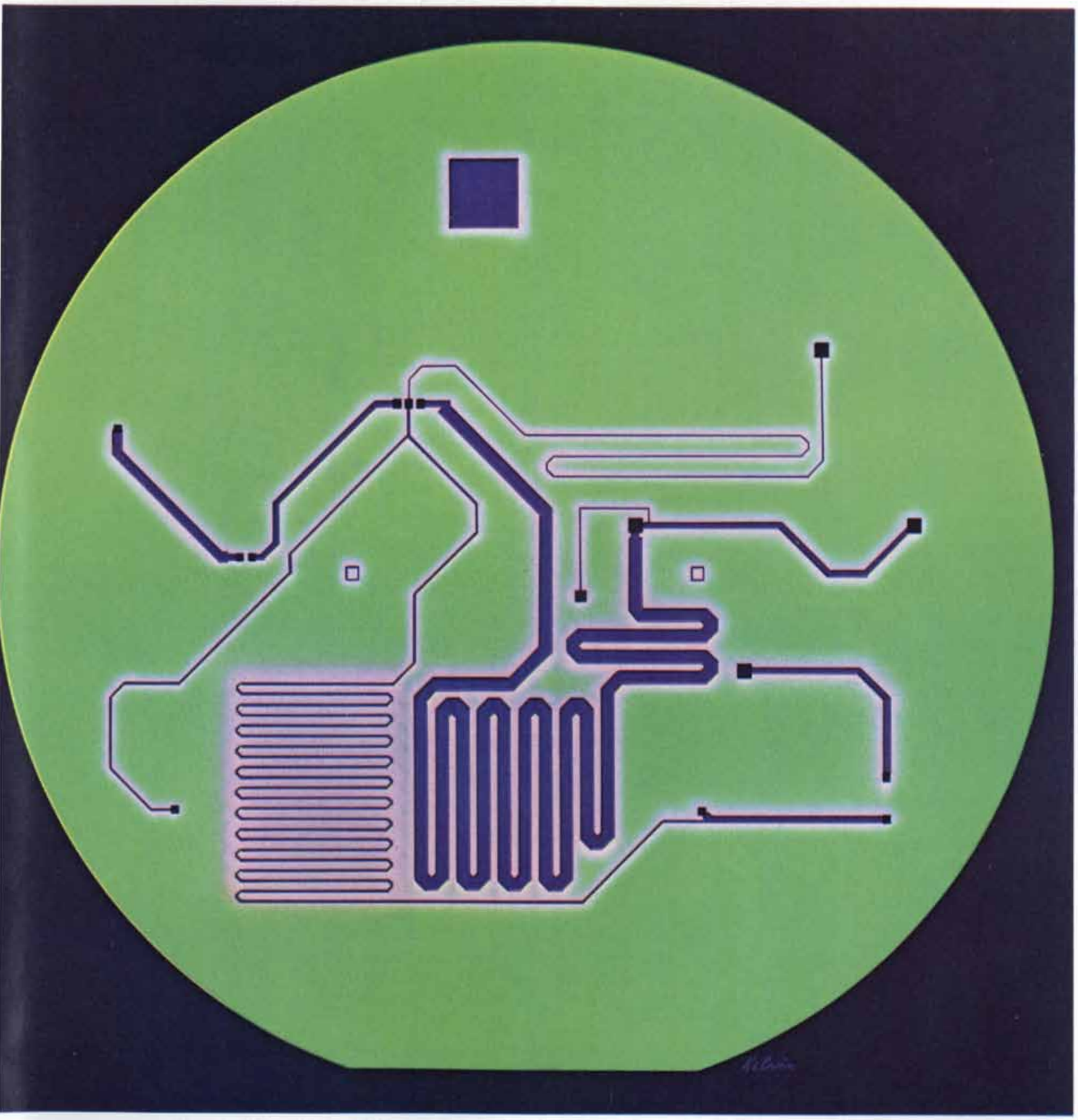
SCIENTIFIC AMERICAN



MICROMECHANICAL DEVICES

$2.50

April 1983

The simple luxury of quiet.

In a noisy world, quiet is a luxury. One that comes standard with the Honda Accord LX Hatchback.

Wind and highway noise are reduced to a whisper by aerodynamic styling, triple door seals and a plush new interior.

Handling is remarkably smooth, the result of more Accord LX standard features. Like variable-assist power steering. Michelin steel-belted radial tires. Front-wheel drive. And 4-wheel independent suspension.

Even air conditioning is standard. Simply push a button and cool air flows through Honda's full-width instrument panel vents.

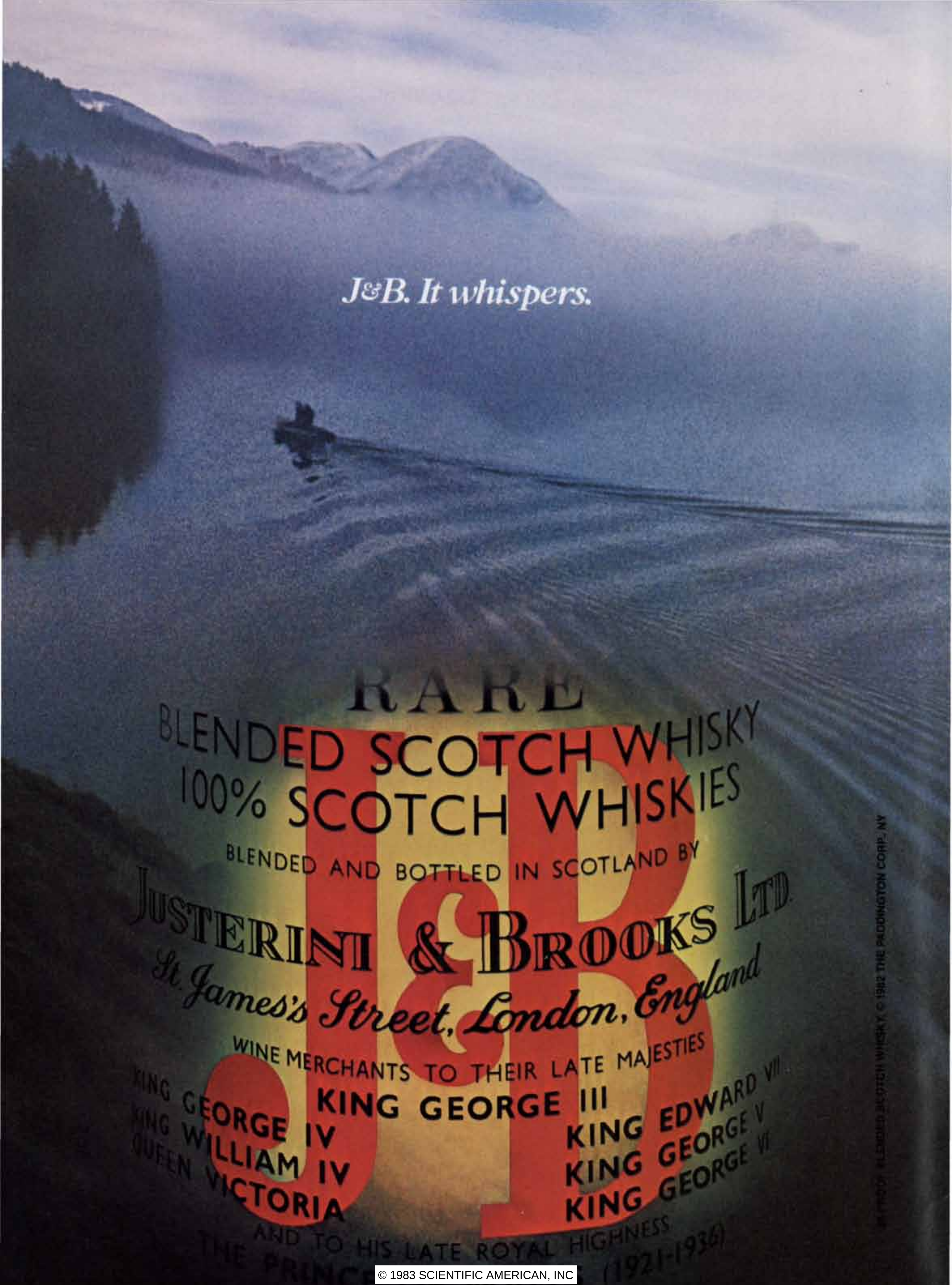
The instrument panel has its own quiet good looks. Tachometer, maintenance reminder, electronic warning system, quartz digital clock and calendar are all sensibly arranged.

The AM/FM stereo with cassette player is optional. And you can also choose a new 4-speed automatic transmission that makes this lap of luxury a joy to drive.

One ride in the Accord LX Hatchback is simple proof. Luxury should be seen and not heard.

HONDA

We make it simple.


J&B. It whispers.
RARE
BLENDED SCOTCH WHISKY
100% SCOTCH WHISKIES
BLENDED AND BOTTLED IN SCOTLAND BY
JUSTERINI & BROOKS LTD.
St. James's Street, London, England
WINE MERCHANTS TO THEIR LATE MAJESTIES
KING GEORGE III
KING GEORGE IV
KING WILLIAM IV
QUEEN VICTORIA
KING EDWARD VII
KING GEORGE V
KING GEORGE VI
AND TO HIS LATE ROYAL HIGHNESS
(1921-1936)
© 1982 THE PADDINGTON CORP., NY

Established 1845 SCIENTIFIC AMERICAN April 1983 Volume 248 Number 4

ARTICLES

DEPARTMENTS

SCIENTIFIC AMERICAN (ISSN 0036-8733), PUBLISHED MONTHLY BY SCIENTIFIC AMERICAN, INC., 415 MADISON AVENUE, NEW YORK, N.Y. 10017. COPYRIGHT © 1983 BY SCIENTIFIC AMERICAN, INC. ALL RIGHTS RESERVED. PRINTED IN THE U.S.A. NO PART OF THIS ISSUE MAY BE REPRODUCED BY ANY MECHANICAL, PHOTOGRAPHIC OR ELECTRONIC PROCESS, OR IN THE FORM OF A PHONOGRAPHIC RECORDING, NOR MAY IT BE STORED IN A RETRIEVAL SYSTEM, TRANSMITTED OR OTHERWISE COPIED FOR PUBLIC OR PRIVATE USE WITHOUT WRITTEN PERMISSION OF THE PUBLISHER. SECOND-CLASS POSTAGE PAID AT NEW YORK, N.Y., AND AT ADDITIONAL MAILING OFFICES. AUTHORIZED AS SECOND-CLASS MAIL BY THE POST OFFICE DEPARTMENT, OTTAWA, CANADA, AND FOR PAYMENT OF POSTAGE IN CASH. SUBSCRIPTION RATE: $21 PER YEAR IN THE U.S. AND ITS POSSESSIONS; $27 PER YEAR IN ALL OTHER COUNTRIES. POSTMASTER: SEND ADDRESS CHANGES TO SCIENTIFIC AMERICAN, 415 MADISON AVENUE, NEW YORK, N.Y. 10017.

SCIENTIFIC AMERICAN

CORRESPONDENCE

Offprints of more than 1,000 selected articles from earlier issues of this magazine, listed in an annual catalogue, are available at 75 cents each. Correspondence, orders and requests for the catalogue should be addressed to W. H. Freeman and Company, 4419 West 1980 South, Salt Lake City, UT 84121. Offprints adopted for classroom use may be ordered direct or through a college bookstore. Sets of 10 or more Offprints are collated by the publisher and are delivered as sets to bookstores.

Photocopying rights are hereby granted by Scientific American, Inc., to libraries and others registered with the Copyright Clearance Center (CCC) to photocopy articles in this issue of SCIENTIFIC AMERICAN for the flat fee of $2 per copy of each article or any part thereof. Such clearance does not extend to the photocopying of articles for promotion or other commercial purposes. Correspondence and payment should be addressed to Copyright Clearance Center, Inc., 21 Congress Street, Salem, MA 01970. Specify CCC Reference Number ISSN 0036–8733/83. $2.00 + 0.00.

Editorial correspondence should be addressed to The Editors, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017. Manuscripts are submitted at the authors' risk and will not be returned unless they are accompanied by postage.

Advertising correspondence should be addressed to C. John Kirby, Advertising Director, SCIENTIFIC AMERICAN, 415 Madison Avenue, New York, NY 10017.

Subscription correspondence should be addressed to Subscription Manager, SCIENTIFIC AMERICAN, P.O. Box 5969, New York, NY 10017. The date of the last issue on your subscription is shown in the upper right-hand corner of each month's mailing label. For change of address notify us at least four weeks in advance. Please send your old address (if convenient, on a mailing label of a recent issue) as well as the new one.

Name ______________________

New Address

Street ______________________

City ______________________

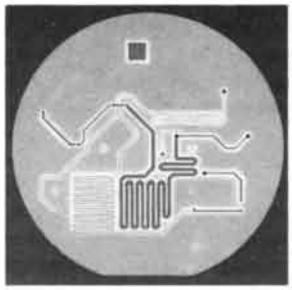
State and ZIP ______________________

Old Address

Street ______________________

City ______________________

State and ZIP ______________________



THE COVER

The painting on the cover illustrates a new technology based on "machining" silicon wafers like those used to make microelectronic chips (see "Silicon Micromechanical Devices," by James B. Angell, Stephen C. Terry and Phillip W. Barth, page 44). The cover painting shows a three-inch wafer that has been machined to make a gas chromatograph that will fit into the palm of the hand. The serpentine elements are gas channels formed by etching grooves into the silicon and bonding the wafer to a glass plate. Square holes etched completely through the wafer allow the gas to flow from system components etched into the front of the wafer to those mounted on the back. Underlying the double and triple sets of holes are valves; the valve seats are etched into the back of the wafer. The sample to be analyzed in the device is injected through a valve into the widest serpentine channel, where it is compressed by a tiny piston. The pressure in the channel is monitored by a silicon sensor mounted on the wafer. On command the pressurized sample is injected into both a lined separation column (such as a hollow fiber of fused silica) and an unlined reference column (the serpentine channel of intermediate diameter on the front of the wafer). Both the separation and the reference columns lead to a gas channel etched into the back of the wafer, over which is mounted a thermal-conductivity detector on a silicon chip. A portable gas analyzer consisting of five such chromatographs and a microcomputer is manufactured by Microsensor Technology, Inc.

THE ILLUSTRATIONS

Cover painting by George V. Kelvin

Page	Source
14–20	Edward Bell
38–39	Ilil Arbel
40–41	U.S. Army
42–43	Sovfoto
45	Stanford University Integrated Circuits Laboratory, courtesy of Phillip W. Barth
46–49	Gabor Kiss
50	Stanford University Integrated Circuits Laboratory, courtesy of Rosemarie Koch
51–52	Gabor Kiss
53–54	Stanford University Integrated Circuits Laboratory, courtesy of Phillip W. Barth (*top*); Gabor Kiss (*bottom*)
55	IBM San Jose Research Laboratory, courtesy of Kurt E. Petersen
56–68	Jerome Kuhl
79–81	Woods Hole Oceanographic Institution
82–92	Ian Worpole
95–96	Tom Prentiss
97	Tom Prentiss (*left*), Adrian R. Morrison (*right*)
98	Adrian R. Morrison
99–102	Tom Prentiss
105	Ryszard Grygiel
106	Ryszard Grygiel (*top*), Peter Bogucki (*bottom*)
107–112	Alan D. Iselin
114–116	Patricia J. Wynne
117	Kenneth B. Raper
118	Patricia J. Wynne (*top*), John Tyler Bonner (*bottom*)
119	Patricia J. Wynne
120	Kenneth B. Raper
123	Michael Goodman
124	Fowler Architectural Collection, Special Collections Division, Milton S. Eisenhower Library, Johns Hopkins University
125–126	Michael Goodman
127	Biblioteca Nazionale Centrale, Florence
128–130	Michael Goodman
133	Matthew G. Wheeler
134–136	Michael Goodman

Pan Am's Clipper Class Seat 18½″

TWA's Ambassador Class Seat 20⅞″

British Airways Super Club Seat 24″*

THE WORLD'S WIDEST AIRLINE SEAT.

Who else but British Airways.

Who else indeed! After all, no other airline on earth offers you a business class in a class with British Airways' Super Club® The world's roomiest business class seat, always next to an aisle or a window, and service as only the British can provide it.

Should you choose to step up a bit, you'll find our Crown First Class™ waiting, with a sleeperseat of noble proportions. It seems a shame to sleep through our haute cuisine and fine china service, but it's nice to be able to.

Finally, there's the stratosphere of luxury, our Concorde. The only supersonic flights to Britain.

And, you'll be pleased to note that your flight miles between the US and London on British Airways will count as credit toward the A Advantage℠ travel award plan.

But, it's no more than you'd exnect from the airline that flies more people to more countries than any other airline. Call your travel agent or British Airways.

British airways

*British Airways Super Club has a few seats 22″ wide because of structural requirements, however, all Super Club Seats are wider than our competitors'.

LETTERS

Sirs:

I was very interested to read Douglas R. Hofstadter's article on nonsense ["Metamagical Themas," SCIENTIFIC AMERICAN, December, 1982], both as a professional mathematician and as a connoisseur of nonsense verse myself. I feel, however, that he omitted two further significant sources of nonsense.

One is the works of the English playwright Tom Stoppard. For example, in *Dogg's Hamlet—Cahoot's Macbeth* (in a context that I cannot fully describe here) Stoppard is rendering part of Act V, scene 8 of *Macbeth* into "Dogg" language thus:

"Rafters Birnam cakehops hobble Dunsinane, fry counterpane nit crossly window-framed, fancifully oblong! Sundry cobbles rattling up so chamberlain. Frantic, Macduff! Fry butter ban loss underlay—November glove!"

Secondly, I draw your attention to computer-generated nonsense. I enclose a sample of my own production. The sonnet "How Can the Purple Yeti Be So Red?" gained some fame in this country when it appeared in *The Times* as part of a discussion on whether computers could write verse. I think computers have a useful role to play here, both in randomly juxtaposing unlikely words and in generating new ones.

How can the purple yeti be so red,
Or chestnuts, like a widgeon,
calmly groan?
No sheep is quite as crooked as a bed,
Though chickens ever try
to hide a bone.
I grieve that greasy turnips
slowly march:
Indeed, inflated is the icy pig:
For as the alligator strikes the larch,
So sighs the grazing goldfish for a wig.
Oh, has the pilchard argued with a top?
Say never that the parsnip is too weird!
I tell thee that a wolf-man will not hop
And no man ever praised the convex
beard.
Effulgent is the day when bishops turn:
So let not then the doctor wake the urn!

Oh anaconda, tell me why the crane
Should be samoan when the toadstools
scream:
A wailing hermit never maims a brain,
Although 'tis true that felons harm
a bream.
My heart is spacious, likewise is it red,
When e'er I see the crazy carrots write;
I lost the briny princess—for a bed
Had madly spluttered as it chewed a
light.
Alas! the days of midwife, elk and bat
Are gone, and now the hungry
bailiffs blink;
Momentous was the crocus, now so fat
And ospreys cannot squash the
smiling drink.
I shall no longer hide the ancient goose:
Life's not an ogre, but a gruesome
moose!

J. R. PARTINGTON

Pembroke College
Cambridge, England

Sirs:

I read Jearl Walker's "Amateur Scientist" column in January with particular interest because I have a special interest in historical measurements of the universal constant of gravitation. There are two details in his résumé of Henry Cavendish's experiment that I should like to correct.

In the first place the torsion-balance experiment was planned by the Reverend John Michell, who unfortunately died before he could use the apparatus he had developed. It was this apparatus that Henry Cavendish took over, and although he reconstructed it, he kept close to Michell's plan and dimensions.

Less recherché, and more important, is the fact that Cavendish did not use an optical lever ("a light beam reflected from a small mirror") to measure the angle of rotation. Instead he observed vernier scales fitted to the ends of the torsion rod that moved against fixed ivory scales. These scales were illuminated by light focused on them from lamps fixed in wall apertures, and telescopes similarly fixed were used to view the scales.

G. S. LEADSTONE

Atlantic College
Glamorgan, South Wales

Sirs:

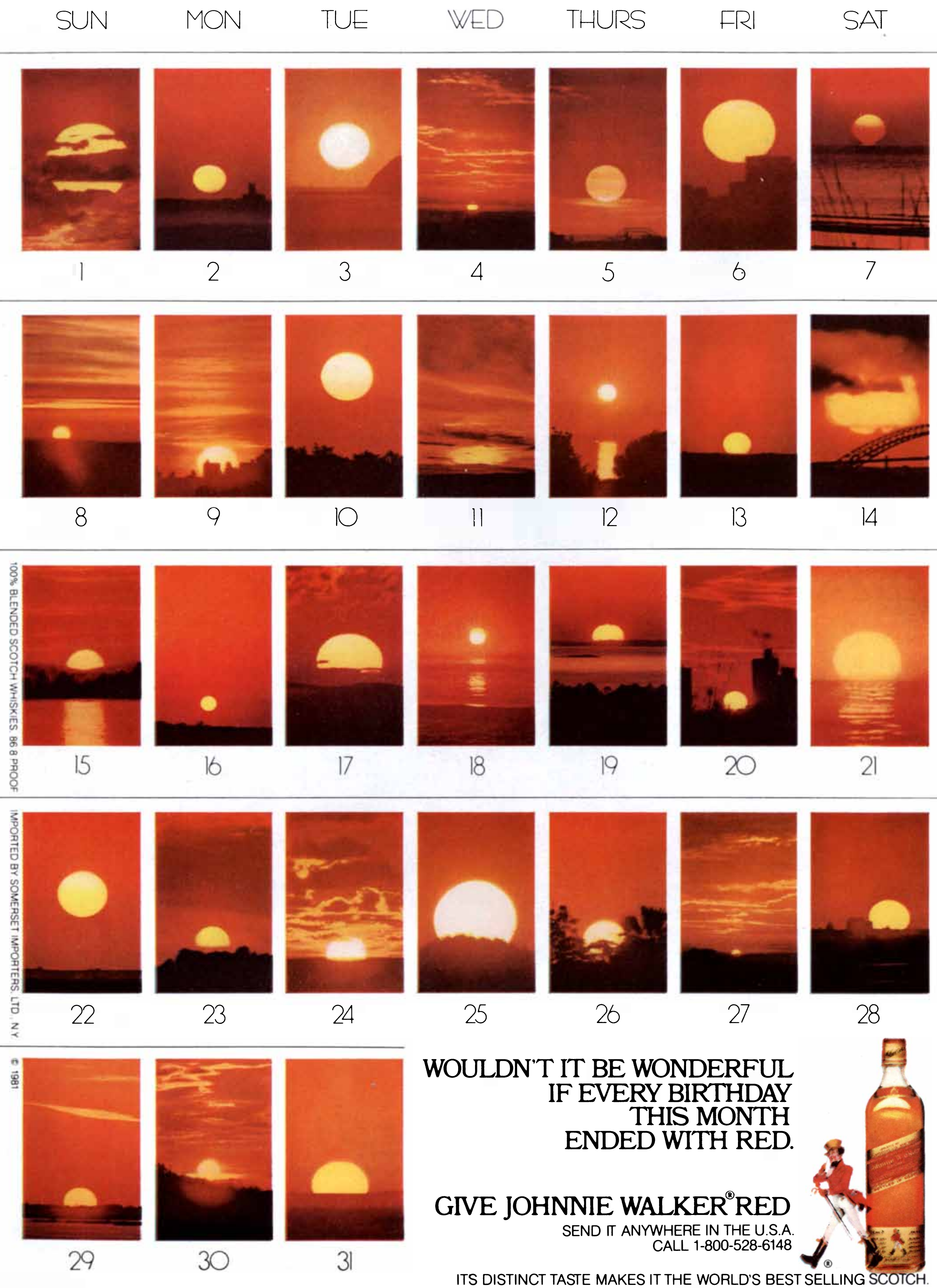
"The Physics of Organ Pipes," by Neville H. Fletcher and Suszanne Thwaites [SCIENTIFIC AMERICAN, January], mentions the "tin-rich alloy" used for diapasons. The late organist and choirmaster Edward Whitney Flint of the Brooks School in North Andover, Mass., told the story that during the London blitz of World War II, the organ and the stained-glass windows of St. Paul's Cathedral were dismantled for safekeeping. After the war the organ was reassembled, except for a few missing pipes. One day the organist happened to be in the upper reaches of the cathedral when he came on a workman. On the scaffold beside him, being melted into solder for replacing the windows, was one of the missing pipes.

SAMUEL R. PHILLIPS

Portola Valley, Calif.


COMPARED TO THE TOP THREE IMPORT SEDANS,
CHEVY CAVALIER PUTS YOU IN THE DRIVER'S SEAT. IN ROOM, POWER AND CHOICE.
*Limited production available after May 1, 1983.
Some Chevrolets are equipped with engines produced by other GM divisions, subsidiaries, or affiliated companies worldwide. See your dealer for details.
USA-1 IS TAKING CHARGE
CAVALIER · CELEBRITY · CITATION · CHEVETTE · CAMARO · MALIBU · MONTE CARLO · CAPRICE · CORVETTE


MORE POWER THAN THE TOP THREE IMPORTS.
Cavalier's new high-compression 2.0 Liter engine outpowers the three top-selling imports (based on SAE Net rated horsepower).
Coupled with Cavalier's front-wheel drive and new available 5-speed, it's performance you can feel.
Plus Electronic Fuel Injection. Chevy Cavalier has it. Accord doesn't. Corolla doesn't and Sentra doesn't.
MORE ROOM THAN THE TOP THREE IMPORT SEDANS.
The Cavalier Sedan shown below has 14 cu. ft. more total room than Toyota Corolla, 10 more than Nissan Sentra, 8 more than Honda Accord (based on EPA ratings for 1983 models).
Cavalier Coupe equals or beats Corolla, Sentra and Honda Prelude coupes in room. And Cavalier Wagon beats the top-selling import wagon by a comfortable margin–11 cu. ft. more room for passengers, 5 more for cargo.
NEW LOWER PRICES FOR 1983.
That's right. Based on a comparison of Manufacturer's Suggested Retail Prices for 1982 and 1983 models, Cavalier's prices are actually lower than last year. Level of equipment will vary.
AND YOU CAN CHOOSE FROM 5 MODELS--INCLUDING A NEW CONVERTIBLE.*
Take your choice—there's room for five in the 4-Door Sedan, 2-Door Coupe, Hatchback, Wagon and room for four in the exciting new Cavalier Convertible.
See your Chevy dealer now. Drive the proof that America's sales leader is taking charge.
Let's get it together... buckle up.
GM
Chevrolet


SUN MON TUE WED THURS FRI SAT
1 2 3 4 5 6 7
8 9 10 11 12 13 14
15 16 17 18 19 20 21
22 23 24 25 26 27 28
29 30 31
WOULDN'T IT BE WONDERFUL
IF EVERY BIRTHDAY
THIS MONTH
ENDED WITH RED.
GIVE JOHNNIE WALKER® RED
SEND IT ANYWHERE IN THE U.S.A.
CALL 1-800-528-6148
ITS DISTINCT TASTE MAKES IT THE WORLD'S BEST SELLING SCOTCH.
100% BLENDED SCOTCH WHISKIES. 86.8 PROOF
IMPORTED BY SOMERSET IMPORTERS, LTD., N.Y.
© 1981

50 AND 100 YEARS AGO

SCIENTIFIC AMERICAN

APRIL, 1933: "The Italian scientist General G. A. Crocco suggests that for flight in the stratosphere at speeds above the speed of sound the internal-combustion engine must be replaced by a totally different type of prime mover, namely the jet-propulsion engine. He revives a suggestion of René Lorin, a Frenchman, who in 1913 proposed an engine that would take in the oncoming air, perhaps at a speed of 1,500 feet per second, through a front opening. The air would be highly compressed on entrance into the engine because of its great speed. Its temperature would rise, and at the same time it would mix with the fuel at 'auto-ignition' temperature; there would be combustion, a further increase in temperature and pressure, and a rush of air out at the rear of the engine at a speed greater than that of the aircraft as a whole. At high altitudes and at speeds of 1,000 miles per hour Lorin's mechanism can probably be made to give an efficiency just as high as that of the present-day combination of a propeller and an internal-combustion engine."

"The fundamental ideas of the behavioristic and Gestalt psychologies justify attempts to construct and develop machines of a new type—machines that can think. Such machines are entirely different from the integraphs, tide calculators and the like to which the term 'thinking machine' is sometimes applied. Unlike the latter they are not designed to perform with mathematical regularity but can 'learn' to vary their actions under certain conditions. Several mechanisms have recently been designed and built to illustrate the 'conditioned reflex,' regarded by behaviorists as the basic element of mental activity. These mechanisms exhibit in simplified form the type of activity studied by the Russian psychologist and physiologist Pavlov in his well-known experiments with dogs. Since it is possible to parallel the conditioned reflex as it is observed in the laboratory, the idea suggests itself that machines embodying this principle can be made to move about and to learn from stimuli encountered during their movement, just as animals do."

"The 'model' stars that astrophysicists use in their investigations must be imaginary—purely mathematical abstractions. We cannot get a sample of stellar material under stellar conditions. We might approximate the correct chemical composition, but pressures of billions of atmospheres and temperatures of billions of degrees are beyond our power to imitate. Each investigator has his own pet assumptions, and this is fortunate, for among the half dozen or so now active we may hope for a good reconnaissance of the possibilities. No one who is not eager for masses of computation should enter this field, for the calculations are laborious in the extreme. The stellar models that have so far been computed are curiously different. Some have central cores so dense that the ordinary gas laws fail and other equations must be used. Some are of enormous size; others are tiny—no larger than some of the planets. When the devoted labor of the computers has given us a large enough choice among them, we may hope to know more than we do now about the actual, as well as the possible, constitution of the stars."

SCIENTIFIC AMERICAN

APRIL, 1883: "By means of two dynamo-electric machines (Gramme system) of identically similar make M. Marcel Deprez has transmitted to Munich (over a distance of 57 kilometers) along an iron telegraph wire of 4.5 millimeters' diameter the power obtained at Miesbach from a steam engine. The receiving machine placed in the Crystal Palace at Munich supplied motive power for eight days to a centrifugal pump feeding a small waterfall about 2.5 meters in height. The success of the transmission of power from Miesbach to Munich is in every way an important event in the history of the technical applications of electricity."

"Dr. Roswell Park of Chicago describes the most recent applications of the electric light for surgical purposes. One form of these new instruments, called the gastroscope, is an instrument for the examination of the stomach. It consists of a bent tube with a window at one end, electric wires and tubes for the introduction of a water circulation to keep the tube cool while the electric light is burning. The tube is also equipped with reflector prisms and lenses for directing the light through it. A variety of other instruments have been made. For example, we have the laryngoscope, for examination of all parts of the throat; the oesophagoscope, for the gullet; the otoscope, for the ears; the urethroscope, for the bladder; the cystoscope, etc. They promise to be of utility and importance for the medical profession, for by their use many parts of the human system heretofore hidden from the eye may now be brilliantly lighted and examined and their condition in disease and health ascertained."

"Is there more than one force? We speak of the forces of nature and classify them as heat force, light force, electric force, etc. Should we not speak of the force of nature as exhibited in heat, in light, in electricity, etc.? You ask, if there is but one force, why does it not always manifest itself in the same way? Simply because the conditions are different at different times and in different places. The convertibility of these expressions of force, the one into the other, is perhaps the strongest proof of their identity of origin. It is impossible to study closely the properties of matter and the various phenomena exhibited in its relation to energy without coming to the conclusion that all forms of energy, however they may appear to our senses, are the offspring of one parent."

"Scientific prophets have foretold that a day will come when the 'residual products' resulting from distilling coal tar will be so valuable as to reduce the price of the gas obtained from the original coal to a mere nothing. A gentleman in England claims to know of an oven for making coke that enables those who use it to drive steam engines without any expense for fuel. Every ton of coal consumed in the oven yields coke worth 7s. and tar and ammonia worth 4s., in addition to 14,000 cubic feet of gas. If, therefore, the first two products are sold, the price—11s.—more than pays for the coal from which they were derived, as well as for labor, wear and tear, and interest on the capital sunk in plant. It is certainly a bold claim to put forward, but it is justified by the present prices of coke, tar and ammonia. If these ovens come into general use, the market value of such products will fall heavily in proportion to the immense enhancement of supply, and the prices fetched would not cover the cost of materials and labor."

"Running as an exercise can strengthen the limbs, develop the lungs, exercise the will and promote the circulation of the blood. The clothing should be light, the head bare and the neck uncovered. Care must be taken not to overdo. Running is well adapted to both young and middle-aged persons but not to those who are fat. Sedentary persons may find great benefit in it after the day's work is ended. If they live in cities, a quiet spot in the park may be selected and short trials adapted to the strength entered into. Girls can run as well as boys, and while they cannot go so fast, they can run much more gracefully. After puberty the change in the formation of the bones of the pelvis in girls renders running less easy and graceful. The modern style of dress for girls after puberty is also entirely unsuited to running."

THE AUTHORS

BARRY M. BLECHMAN and MARK R. MOORE ("A Nuclear-Weapon-Free Zone in Europe") are associated with the Roosevelt Center for American Policy Studies in Washington. Blechman is vice-president of the Roosevelt Center, responsible for its activities on foreign policy and defense policy. He received his undergraduate education at Queens College of the City University of New York. He then went on to obtain his master's degree in 1964 at New York University and his doctorate in international relations in 1971 from Georgetown University. From 1966 to 1971 he was on the staff of the Center for Naval Analyses, a research organization sponsored by the Federal Government. In 1971 he moved to the Brookings Institution, taking a leave in 1976 to become a member of President Carter's transition planning staff at the Office of Management and Budget. The following year he was appointed assistant director of the U.S. Arms Control and Disarmament Agency. He left the Government in 1979 to become a senior associate of the Carnegie Endowment for International Peace, moving from there to the Roosevelt Center. Blechman was a staff expert for the International Commission on Disarmament and Security Issues and helped to write its final report, in which the plan for a battlefield-nuclear-weapon-free zone in Europe was originally put forward. Moore is a graduate student at the Johns Hopkins School of Advanced International Studies and research assistant at the Roosevelt Center. He was graduated from Amherst College with a B.A. in 1981.

JAMES B. ANGELL, STEPHEN C. TERRY and PHILLIP W. BARTH ("Silicon Micromechanical Devices") are electrical engineers who did the work that forms the basis of their article when all three were at Stanford University. Angell is professor of electrical engineering and associate chairman of the department of electrical engineering at Stanford. He has three degrees from the Massachusetts Institute of Technology: an S.B. and an S.M. (both given in 1946) and an Sc.D. in electrical engineering (1952). From 1951 to 1960 he was a member of the research division of the Philco Corporation. At the end of that time he went to Stanford. An enthusiastic amateur musician, Angell is university carillonneur at Stanford, where he plays the 35-bell carillon in Hoover Tower. Terry is the founder of Microsensor Technology, Inc. His Ph.D. in electrical engineering (1975) is from Stanford, where he was senior research associate until he left in 1981 to found his company. Barth is senior research associate at the Stanford Integrated Circuits Laboratory. He was graduated from the University of Notre Dame with a B.S. in 1973 and went on to earn his M.S. (1974) and his Ph.D. in electrical engineering (1980) from Stanford. He is currently working with Angell on biomedical applications of miniature silicon sensors.

HAIM HARARI ("The Structure of Quarks and Leptons") is Annenberg Professor of High Energy Physics at the Weizmann Institute of Science. A fifth-generation Israeli, he received his master's degree and his doctorate in physics from the Hebrew University in Jerusalem. After serving for four years in the Israeli Army he joined the faculty of the Weizmann Institute in 1966. He is chairman of the planning and grants committee of the Council of Higher Education, the institution responsible for administering government financial support to the universities of Israel.

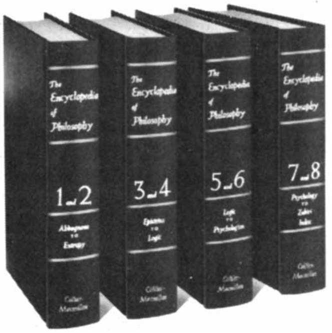
JOHN M. EDMOND and KAREN VON DAMM ("Hot Springs on the Ocean Floor") are marine geochemists. Edmond is professor of marine geochemistry at the Massachusetts Institute of Technology. He writes: "I am a native of Glasgow and got a degree in chemistry from the university there. Then I went to the Scripps Institution of Oceanography at the University of California at San Diego and did a Ph.D. on the carbon dioxide system in seawater. Following that I joined the faculty at M.I.T. The research efforts of my group there are divided about equally among three areas: the trace-metal geochemistry of seawater and general marine geochemistry; the chemistry of ridge-crest hot springs, and chemical processes in large lakes and big rivers and their estuaries. My outside interests include the politics, history and cultures of the overseas countries in which I have the opportunity to do field work. My current interest is in the linguistic development of my two-year-old Scots-Iranian son." Von Damm is a graduate student in the Joint Program in Oceanography conducted by M.I.T. and the Woods Hole Oceanographic Institution. She is finishing her doctoral thesis on the detailed chemical analysis of the waters vented by hot springs on the ocean floor.

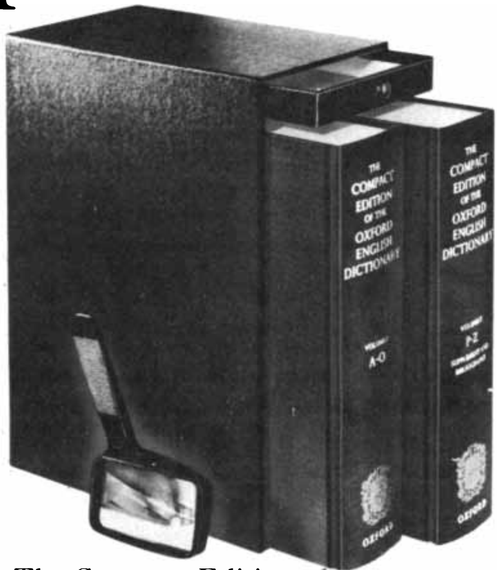
ADRIAN R. MORRISON ("A Window on the Sleeping Brain") is professor of anatomy at the School of Veterinary Medicine at the University of Pennsylvania. He has two degrees from Cornell University: a D.V.M. (1960) and an M.S. (1962). His Ph.D. in anatomy (1964) is from the University of Pennsylvania. After a year of postdoctoral work at the University of Pisa he returned to Pennsylvania, where he has remained.

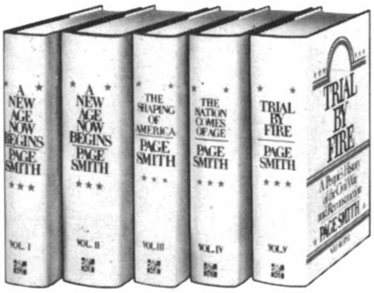
PETER BOGUCKI and RYSZARD GRYGIEL ("Early Farmers of the North European Plain") are archaeologists who have collaborated in work on the earliest farming cultures of north-central Poland since 1976. Bogucki is lecturer in anthropology at the University of Massachusetts at Boston. In the work with Grygiel his particular interest is the study of animal bones at the sites of the farming communities. He obtained his B.A. at the University of Pennsylvania in 1974 and went on to earn his Ph.D. in anthropology from Harvard University in 1981. In addition to his interest in prehistoric Europe he has been studying patterns of animal utilization by the colonial inhabitants of Portsmouth, N.H., and other New England towns. Grygiel is head of the Neolithic department at the Museum of Archaeology and Ethnography in Łódź. He earned his master's degree at Adam Mickiewicz University in Poznań in 1974 and his doctorate in archaeology from the University of Warsaw in 1980. His major scientific interest is the chronology of the early Neolithic communities in north-central Poland, and he has excavated many sites from the early Neolithic period.

JOHN TYLER BONNER ("Chemical Signals of Social Amoebae") is George M. Moffett Professor of Biology and chairman of the department of biology at Princeton University. He has three degrees from Harvard: a B.A. from Harvard College (1941) and an M.A. (1942) and a Ph.D. in biology (1947) from Harvard University. He writes: "In 1939 while an undergraduate I became interested in cellular slime molds (or social amoebae) as organisms ideally suited to experimental studies in developmental biology. It was a case of love at first sight and we have been going steady ever since. After graduate work and a stint in the Air Corps I came to Princeton, where I remain. Besides continuing my experimental work I have written a number of books, which mostly revolve around the general theme of evolution and development from cells to societies." Among Bonner's books are *Cells and Societies* (1955) and *On Development: The Biology of Form* (1977).

MICHAEL McCLOSKEY ("Intuitive Physics") is assistant professor of psychology at Johns Hopkins University. He was graduated from Emory University with a B.A. in 1975 and went on to get his Ph.D. in cognitive psychology from Princeton University in 1978 before moving to Johns Hopkins.







METAMAGICAL THEMAS

In which a discourse on the language Lisp concludes with a gargantuan Italian feast

by Douglas R. Hofstadter

Last month I described Édouard Lucas's Tower of Brahma puzzle, where the object is to transfer a tower of 64 gold disks from one diamond needle to another, making use of a third needle on which disks can be placed temporarily. The disks must be picked up and moved one at a time, the only other constraint being that no disk may ever be placed on a smaller one. The problem I posed for readers was to come up with a recursive description, expressed as a function in the computer language Lisp, of how to accomplish this goal (and thereby, according to the priests of Brahma, end the world).

I pointed out that the recursion is evident enough: in order to transfer 64 disks from one needle to another, using a third, it suffices to know how to transfer 63 disks from one needle to another, using a third. To recapitulate the idea, it is this. Suppose the 64-disk tower of Brahma starts out on needle *a*. At the top in the illustration on this page is a schematic picture representing all 64 disks by a mere 4. First of all, relying on your presumed 63-disk-moving ability, transfer 63 disks from needle *a* to needle *c*, using needle *b* as your "helping needle." How the setup looks is shown second from the top in the illustration. (Note that 4 plays the role of 64 in my original picture, so that 3 plays the role of 63, but for some reason 1 does not play the role of 61. Isn't that peculiar?) Now, simply pick up the one remaining *a* disk—the biggest disk of all—and plunk it down on needle *b*, as is shown third from the top.

```
   o|o          |           |
  oo|oo         |           |
 ooo|ooo        |           |
oooo|oooo       |           |
=========   =========   =========
    a           b           c

    |           |           |
    |           |          o|o
    |           |         oo|oo
oooo|oooo       |        ooo|ooo
=========   =========   =========
    a           b           c

    |           |           |
    |           |          o|o
    |           |         oo|oo
    |      oooo|oooo     ooo|ooo
=========   =========   =========
    a           b           c

               oo
    |           |           |
    |         oo|oo         |
    |        ooo|ooo        |
    |       oooo|oooo       |
=========   =========   =========
    a           b           c
```

The Tower of Brahma puzzle with four disks

Now you can see how easy it will be to finish up: simply reexploit your 63-disk ability to transfer that pile on *c* back to *b*, this time using *a* as the "helping needle." Notice how in this maneuver needle *a* plays the helping role needle *c* played in the preceding 63-disk maneuver. A split second before the last disk is put in place the situation is as it is shown at the bottom in the illustration. Why before it is in place? Why not after? Because the entire world then turns to dust, and it is too hard to draw dust.

Now, someone might complain that I left out all the hard parts: "You just magically assumed an ability to move 63 disks." So it may seem, but there is nothing magical about such an assumption. After all, in order to move 63 you merely need to know how to move 62. And to move 62, you merely need to know how to move 61. On it goes down the line, until you bottom out at the "embryonic case" of the Tower of Brahma puzzle, the 1-disk puzzle. Now, I admit that you have to keep track of where you are in the process, which may be a bit tedious, but it is merely bookkeeping. In principle you now could actually carry out the entire process—if you were bent on seeing the world end!

As our first approximation of a Lisp function let us write an English description of the method. Let us call the three needles "sn", "dn" and "hn", standing for source needle, destination needle and helping needle. Here goes:

```
To move a tower of height n
  from sn to dn, making use of hn:
    if n = 1,
       then just carry that one disk
       directly from sn to dn;
    otherwise,
       do the following three steps:
          (1) move a tower of height
          n − 1 from sn to hn,
             making use of dn;
          (2) carry 1 disk from sn to dn;
          (3) move a tower of height
          n − 1 from hn to dn,
             making use of sn.
```

Here the steps labeled 1 and 3 are the two recursive calls; skirting paradox, they seem to call on the very ability they are helping to define. The saving feature is that they involve n − 1 disks instead of n. Note that in step 1 hn plays the "destination" role while dn plays the "helper" role. And in step 3 hn plays the "source" role while sn plays the "helper" role. Since the whole thing is recursive, every needle will be switching hats many times over in the course of the transfers. That is the beauty of the puzzle and in a way it is also the beauty of recursion.

How do we make the transition from English to Lisp? It is quite simple:

```
(def move-tower
 (lambda (n sn dn hn)
  (cond
   ((eq n 1) (carry-one-disk sn dn))
   (t (move-tower (sub1 n) sn hn dn)
     (carry-one-disk sn dn)
     (move-tower (sub1 n) hn dn sn)))))
```

Where are the Lisp equivalents of the English words "from," "to," and "making use of"? They seem to have disappeared. Then how can the Lisp genie know which needle is to play which role at each stage? The answer is that the information is conveyed *positionally.* There are, in this function definition, four parameters: one integer and three "dummy needles." The first of the three needles is the source, the second the destination and the third the helper. Thus in the initial list of parameters (following the "lambda") they are in the order "sn dn hn". In the first recursive call, the Lisp translation of step 1, they are in the order "sn hn dn", showing how hn and dn have switched hats. In the second recursive call, the Lisp translation of step 3, you can see that hn and sn have switched hats.

The point is that the atom names "sn", "dn" and "hn" carry no intrinsic meaning to the genie. They could as well have been "apple", "banana" and "cherry". Their meanings are defined operationally, by the places where they appear in the various parts of the function defini-

How to turn data you can see.

There are more people in more places doing more things with Apples than with any other personal computer in the world.

Which is saying a lot.

But we'd like to take the time to explain just one of the things that can make an Apple® Personal Computer more meaningful to you, personally.

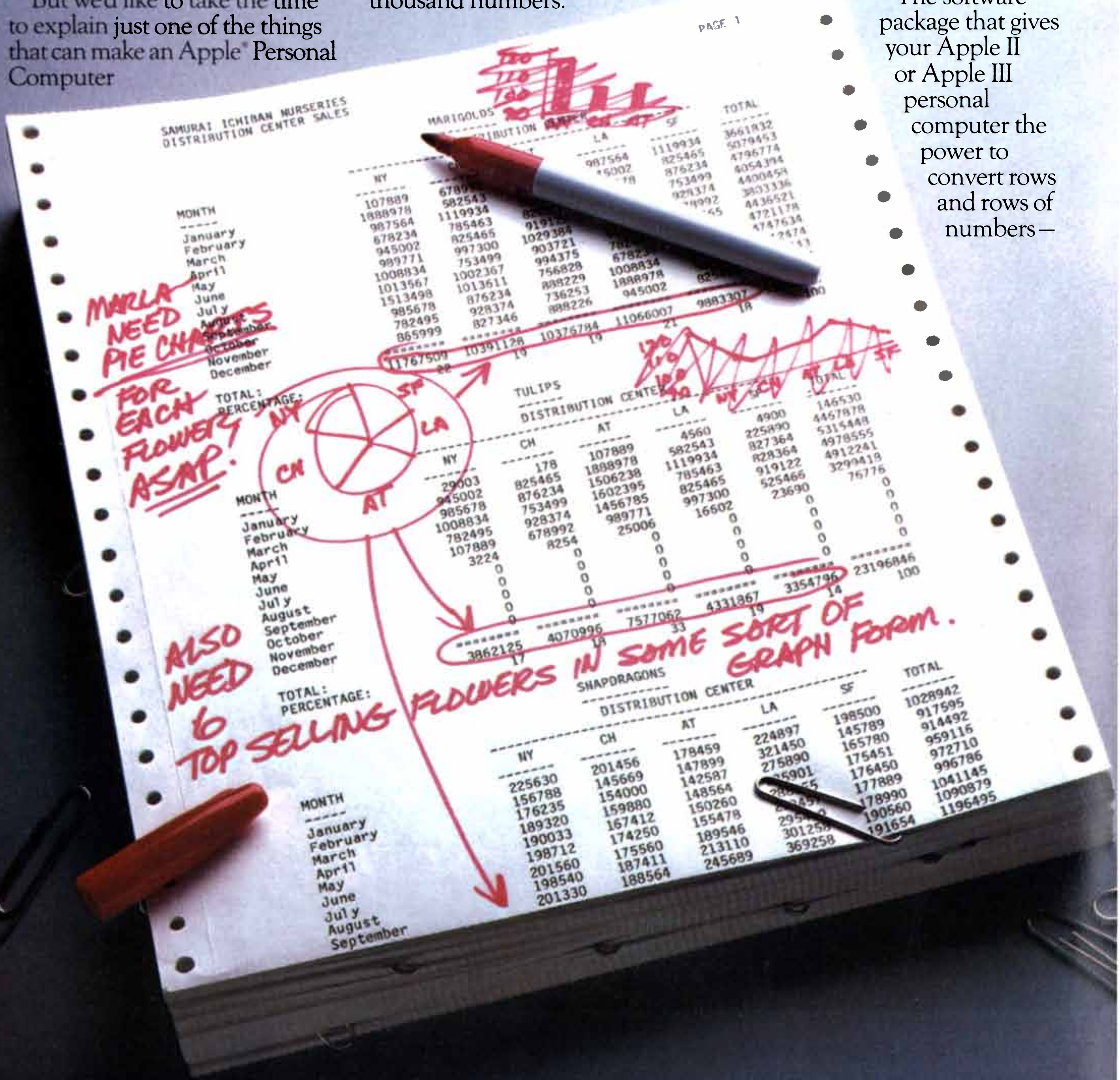
It's called business graphics.

An important business tool that demonstrates, quite graphically, how one picture is worth a few thousand numbers.

More ways to see what you're doing.

That's the beauty of Apple Business Graphics.

The software package that gives your Apple II or Apple III personal computer the power to convert rows and rows of numbers—



a sea of data into

or pounds of printout—into one simple, colorful, comprehensible illustration.

And that gives you the power, for example, to thoroughly understand relationships and comparisons in a complex market analysis. Without having to spend hours sifting through

Horizontal bar graph (a unique feature) makes it easy to compare the profitability of six different products — from periwinkles to petunias.

dull data—armed with little more than a ruler, a compass and a few not-so-magic markers.

With Apple Business Graphics, all you do is enter the data or transfer it directly from your VisiCalc* model or a number of other computer files. Your Apple will take it from there.

Helping you to get the big picture in a variety of ways. Line or area graphs, bar graphs, pie charts or scattergrams.

Whichever configuration best suits your application.

When you're through charting market shares you can do the same with forecasts, budgets, stock trends, business plans or customer demographics.

Show your stuff.

In addition to helping you better visualize results, Apple Business Graphics can make presentations more presentable.

Because whatever you've produced can be rapidly, easily and colorfully reproduced, on virtually any printer or plotter on the market.

Moreover, sophisticated Apple features like exploded views,

To get the overall market share picture, (and any other percentage related data) Ichiban management consults this easy-to-digest pie chart.

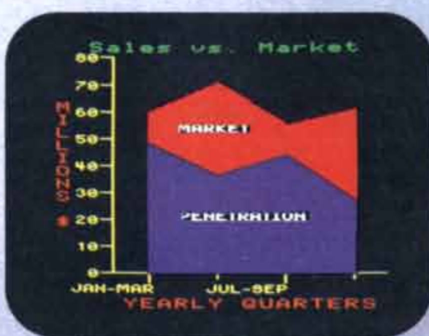
unlimited overlays and floating titles allow your displays to be remarkably flexible.

By combining Apple Business Graphics with various slide-show packages, charts and graphs on your monitor can easily be converted into impressive slides and transparencies.

Getting started in pictures.

Start by seeing any of the over 1300 authorized Apple dealers.

They'll show you Apple Business Graphics software can generate more types of pictures, in more colors, using more data



With up to six colors per graph, employees can easily see just how their annual sales blossomed relative to the rest of the market.

than any other graphics package.

More important, they'll show you what all that can mean to you, your accounting firm, your import business, or your chain of quiche-on-a-bun family restaurants.

They'll even help you produce your first picture.

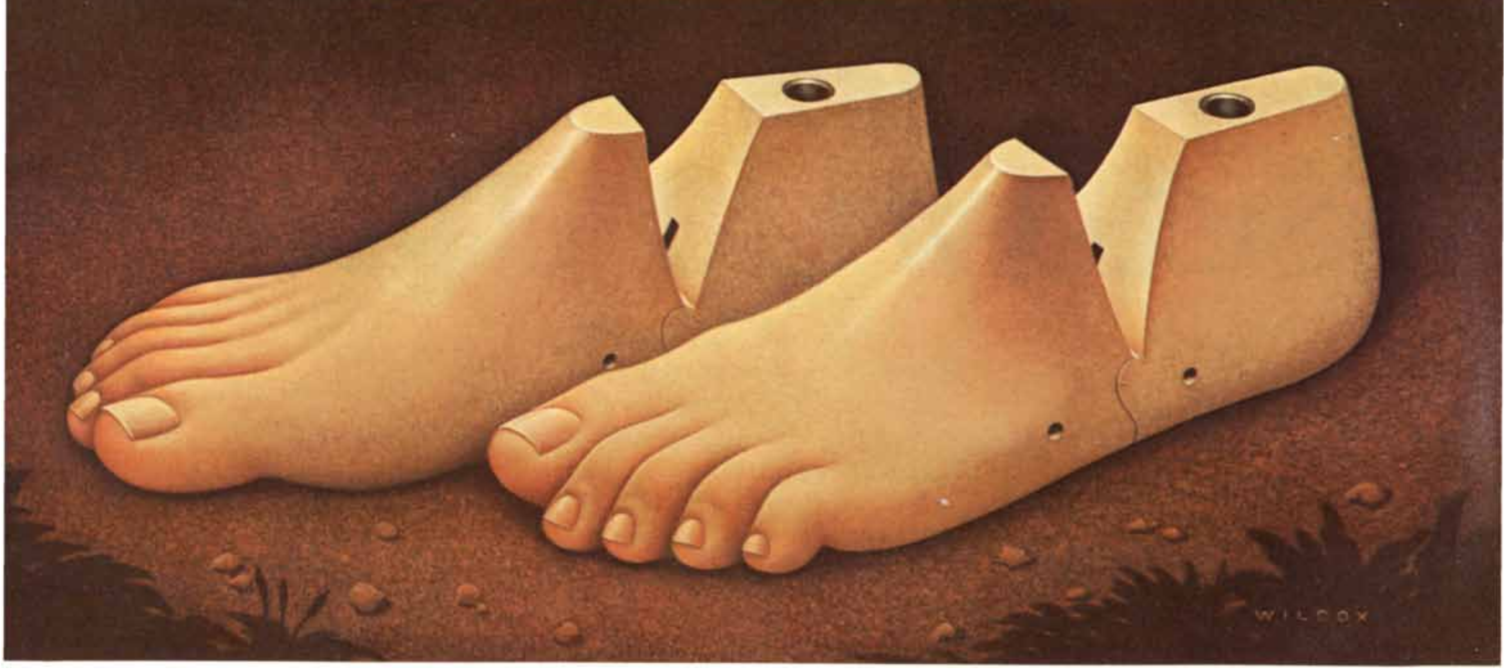
But business graphics is just one of the things an Apple can help you do. One of thousands. Which is not terribly surprising, when you consider there's more software available for Apples than for any other personal computer.

So consider a trip to your Apple dealer. And draw your own conclusions.

apple

The most personal computer.

Call (800) 538-9696 for the location of the authorized Apple dealer nearest you, or for information regarding corporate purchases through our National Account Program. In Calif. (800) 662-9238. Or write Apple Computer Inc., Advertising and Promotion Dept., 20525 Mariani Ave., Cupertino, CA 95014. VisiCalc is a registered trademark of VisiCorp.



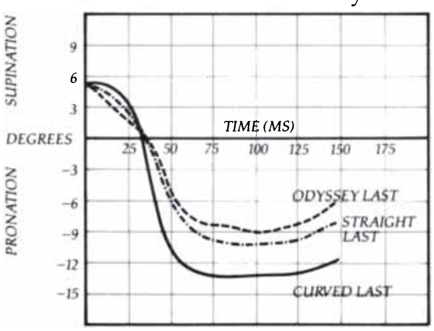
A NEW SHOE WILL NEVER FIT BETTER THAN THE LAST.

Let's get our terms straight.

From now on, when we use the word "last," we're talking about what comes first.

We're talking about that rather strange looking device over which the shoe is initially formed. Now this last can be made of wood, metal or plastic. But what's truly critical, what keeps us up nights, is its shape.

In fact, at Nike, we think the shape of the last is so important that we've launched one of the most comprehensive foot morphology studies since the 1940's. Because the only way to give you a better fitting shoe is to come up with a last that more accurately



A different last means different degrees of rearfoot control. Data obtained from high speed films of 10 subjects running at 7:00 mile pace in shoes based on three different lasts.

reflects both the structure and dynamics of your individual foot.

It's a tough job. And we're far from done. But with the data we've already collected, combined with extensive fit trials and wear tests, we've been able to make some radical changes. Changes that not only affect the comfort of Nike shoes, but more surprisingly, also improve their inherent stability.

The normal foot, as seen in each of three new Nike lasts. The Odyssey provides the most support under the medial arch.

Today, there are three new Nike lasts. With each, we widened the heel base, trimmed the Achilles area and put what amounts to a slight curve in the heel area. As a result, the foot is now centered more directly over the middle of the shoe.

Which, as tests prove, effectively reduces rearfoot motion.

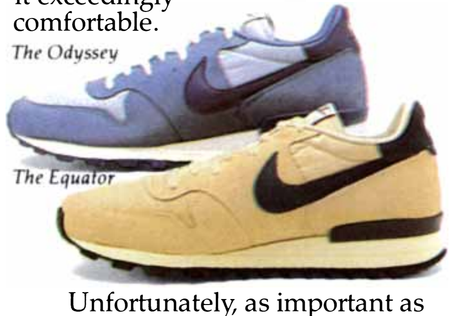
Up front, we gave the big toe additional room. Both overhead and on the medial side. This too promotes stability and may even have a beneficial influence on the foot's ability to create its own cushion.

Our recommendations:

If you're a fast runner or have high-arched rigid feet, try a model from our new curved last. Namely, the Columbia, Aurora, Valkyrie, Internationalist, Pegasus, or the Elite Classic.

On the other hand, if you run a bit slower, or have low-arched flexible feet, or pronate to excess (that is, roll to the inside of your shoes), slip into the new Equator made on our new straight last. It does an even better job of controlling rearfoot motion because of the increased support under the medial arch.

Our most revolutionary last, however, we dubbed the Odyssey. Then we built a shoe of the same name to go with it. As a last, it's the straightest of the bunch. And also the most versatile. Runners who pronate a lot, receive a lot of correction. Those who pronate less, get less. And just about everyone finds it exceedingly comfortable.



Unfortunately, as important as a last may be, you'll probably never see one. They simply never leave our factory.

It's frustrating, we know. But then again, sometimes what helps you the most is the thing that gives you fits.

NIKE
Beaverton, Oregon

```
(ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab)
```

The Lisp genie fires back the moves to solve a 9-high Tower of Brahma puzzle

tion. Hence it would have been a gross blunder to have written, for instance, "(move-tower (sub1 n) sn dn hn)" as Lisp for step 1, because that contains no indication that hn and dn must switch roles in that step.

An important question remains. What happens when the friendly Lisp genie comes to a step that says "carry-one-disk"? Does it suddenly zoom off to the temple at Benares and literally heft a solid gold disk? Or more prosaically, does it pick up a plastic disk on a table and transfer it from one plastic needle to another? In other words, does some physical action, rather than a mere computation, take place?

Well, in theory it is quite possible. In fact, even in practice the execution of a Lisp command could actually cause a mechanical arm to move to a specific location, to pick up whatever its mechanical hand grasps, to carry that object to another specific location and then to release it. In these days of industrial robots there is nothing science-fictional about it. In the absence of a mechanical arm and hand to move physical disks, however, what could it mean?

One obvious and closely related possibility is for there to be a display of the puzzle on a screen, and for "carry-one-disk" to cause a *picture* of a hand to move, to *appear* to grasp a disk, pick it up and replace it somewhere else. That would amount to simulating the puzzle graphically, which can be done with varying degrees of realism.

Suppose, however, we do not have fancy graphics hardware or software at our disposition. Suppose all we want is to create a printed recipe telling us how to move our soft, organic, human hands in order to solve the puzzle. Thus in the case of a three-disk puzzle the desired recipe might read "ab ac bc ab ca cb ab", or equally well "12 13 23 12 31 32 12". What could help us to reach this humbler goal?

If we had a program that moved an arm, we would be concerned not with the value it returned but with the patterned sequence of "side effects" it carried out. Here, in contrast, we are concerned only with the value our program is going to return: a patterned list of atoms. The list for n = 3 has got to be built up from two lists for n = 2. This idea was shown at the end of last month's column:

ab ac bc   ab   ca cb ab

In order to set groups apart in Lisp, rather than using wider spaces we use parentheses. Hence our goal for n = 3 might be to produce the sandwichlike list "((ab ac bc) ab (ca cb ab))". One way to produce a list out of its components is to use "cons" repeatedly. Therefore if the values of the atoms "apple", "banana" and "cherry" are 1, 2 and 3, then the value returned by "(cons apple (cons banana (cons cherry nil)))" will be the list "(1 2 3)". There is, however, a shorter way to get the same result, namely to write "(list apple banana cherry)". It returns the same value. Similarly, if the atoms "sn" and "dn" are bound respectively to "a" and "b", then execution of the command "(list sn dn)" will return "(a b)". The function "list" is an unusual function in that it takes any number of arguments at all—even none, so that "(list)" returns the value of nil!

Now let us tackle the problem of what value we want the function called "carry-one-disk" to return. It has two parameters that represent needles, and ideally we should like it to return a single atom made out of those needles' names, such as "ab" or "12". For the moment it will be easier if we assume that the needle names are the numbers 1, 2 and 3. In this case to make the number 12 out of 1 and 2 it suffices to do a bit of arithmetic: multiply the first number by 10 and add on the second. Here is the Lisp for that:

```
(def carry-one-disk
 (lambda (sn dn)
  (plus (times 10 sn) dn)))
```

On the other hand, if the needle names are non-numeric atoms, we can use a standard Lisp function called "concat", which takes the values of its arguments (any number, as with "list") and concatenates them all to make one big atom. Thus "(concat 'con 'cat 'e 'nate)" returns the atom "concatenate". In such a case we could write:

```
(def carry-one-disk
 (lambda (sn dn) (concat sn dn)))
```

Either way we have solved the bottom-line half of the move-tower problem.

The other half of the problem is what the recursive part of move-tower will return. Well, that is pretty easy. We simply want it to return a sandwichlike list in which the values of the two recursive calls flank the value of the single call to carry-one-disk. And so we can modify our previous recursive definition very slightly, by adding the word "list":

```
(def move-tower
 (lambda (n sn dn hn)
  (cond
   ((eq n 1) (carry-one-disk sn dn))
   (t
    (list
     (move-tower (sub1 n) sn hn dn)
     (carry-one-disk sn dn)
     (move-tower (sub1 n) hn dn sn))))))
```

Now let us conduct a little exchange with the Lisp genie:

```
—> (move-tower 4 'a 'b 'c)
(((ac ab cb) ac (ba bc ac))
              ab
 ((cb ca ba) cb (ac ab cb)))
—>
```

Smashing! It actually works! Isn't that pretty? In last month's column this list was called "brahma".

Suppose we wanted to suppress all the inner parentheses, so that just a long, uninterrupted sequence of atoms would be printed out. For instance, we would get "(ac ab cb ac ba bc ac ab cb ca ba cb ac ab cb)" instead of the intricacy of brahma. It would be slightly less informative but more impressive in its opaqueness. In this case we would not want to use the function "list" to make our sandwich of three values but would have to use some other function that would remove the parentheses from the two flanking recursive values.

This is a case where the standard Lisp function "append" comes in handy. It splices any number of lists together, dropping their outermost parentheses as it does so. Thus "(append '(a (b)) '(c) nil '(d e))" yields the five-element list "(a (b) c d e)" rather than the four-element list "((a (b)) (c) nil (d e))", which would be yielded if "list" rather than "append" had appeared in the function position. Using append and a slightly modified version of carry-one-disk to work with it, we can formulate a final definition of move-tower that does what we want:

```
(def move-tower
 (lambda (n sn dn hn)
  (cond
   ((eq n 1) (carry-one-disk sn dn))
   (t
    (append
     (move-tower (sub1 n) sn hn dn)
     (carry-one-disk sn dn)
     (move-tower (sub1 n) hn dn sn))))))

(def carry-one-disk
 (lambda (sn dn) (list (concat sn dn))))
```

To test this out I asked the Lisp genie to solve a 9-high Tower of Brahma puzzle. What it shot back at me, almost instantaneously, is at the top of the preceding page.

We have just been through a rather sophisticated and brain-taxing example of recursion. Now let us take a look at a recursion that offers us a different kind of challenge. This recursion comes from an offhand remark I made in the last column. I used the odd variable name "tato", mentioning that it is a recursive acronym standing for "tato (and tato only)". Using this fact you can expand "tato" any number of times. The sole principle is that each occurrence of "tato" on a given level is replaced by the two-part phrase "tato (and tato only)" to make the next level. How it goes is shown in the illustration below. For us the challenge is to write a Lisp function that returns the value of "tato" after n recursive expansions, for any n. It is irrelevant that for n much bigger than 3 the whole thing gets ridiculously large. We are theoreticians!

There is only one problem: any Lisp function must return a single Lisp structure (an atom or a list) as its value, and the entries in our table do not satisfy this criterion. For instance, the entry for n = 2 consists of one atom followed by two lists. To fix this we can turn each entry in the table into a list by enclosing it in one outermost pair of parentheses. Now our goal is consistent with Lisp. How do we attain it?

Recursive thinking tells us that the bottom line, or embryonic case, occurs when n equals 0, and that otherwise the nth line is made from the line before it by replacing the atom "tato", wherever it occurs, with the list "(tato (and tato only))", but without its outermost parentheses. We can write this up:

```
(def tato-expansion
 (lambda (n)
  (cond
   ((eq n 0) '(tato))
   (t
    (replace
        'tato
        '(tato (and tato only))
        (tato-expansion (sub1 n)))))))
```

The only thing is that we have not specified what we mean by "replace". We must be very careful in defining how to carry out our "replace" operation. Look at any line of the "tato" table and you will see that it contains one more element than the preceding line. Why is this? Because the atom "tato" gets replaced each time by a two-element list whose parentheses, as I pointed out above, are dropped in the act of replacement. It is this parenthesis dropping that is the sticky point. A less tricky example of such parenthesis-dropping replacement than the recursive one involving "tato" would be "(replace 'a '(1 2 3) '(a b a))", whose value should be "(1 2 3 b 1 2 3)". Rather than exact substitution of a list for an atom, this kind of replacement involves splicing or appending inside a long list.

Let us try to specify in Lisp—using recursion as usual—just what we mean by "replacing" all occurrences of the atom "atm" by a list called "lyst", inside a long list called "longlist". This is a good puzzle for you to try. Let me give you a hint: See how the answer for argument "(a b a)" is built out of the answer for argument "(b a)". Also look at other simple cases like that, moving back down toward the embryonic case.

The embryonic case occurs when longlist is nil. Then nothing happens, and so our answer should be nil.

The recursive case involves building a more complex answer out of a simpler one assumed to be given. We can fall back on our "(a b a)" example for this. We built the complex answer "(1 2 3 b 1 2 3)" out of the simpler answer "(b 1 2 3)" by appending "(1 2 3)" to it. On the other hand, we could consider "(b 1 2 3)" itself to be a complex answer built out of the simpler answer "(1 2 3)" by consing "b" onto it. Why does one involve appending and the other involve consing? Simply because the first case involves the atom "a", which *does* get replaced, whereas the second involves the atom "b", which does *not* get replaced. This observation enables us to attempt to write down a recursive definition of "replace", as follows:

```
(def replace
 (lambda (atm lyst longlist)
  (cond
   ((null longlist) nil)
   ((eq (car longlist) atm)
    (append
     lyst
     (replace atm lyst (cdr longlist))))
   (t
    (cons
     (car longlist)
     (replace atm lyst (cdr longlist)))))
 ))
```

As you can see, there is an embryonic case (where longlist equals nil) and then one recursive case featuring "append" and one recursive case featuring "cons". Now let us try out this definition on a new example:

```
—> (replace 'a '(1 2 3) '(a (a) b a))
(1 2 3 (a) b 1 2 3)
—>
```

Whoops! It *almost* worked, except that one of the occurrences of "a" was completely missed. This means that in our definition of "replace" we must have overlooked some eventuality. Indeed, if you go back, you will see that an unwarranted assumption slipped in right under our nose, namely that the elements of longlist are always atoms. We ignored the possibility that longlist might contain sublists. What to do in such a case? The answer: Do the replacement inside those sublists as well. And inside sublists of sublists too—and so on. Can you figure out a way to fix up the definition?

We have seen a recursion before where all parts on all levels of a structure needed to be explored; it was the function "atomcount" last month, in which we did a simultaneous recursion on both "car" and "cdr". The recursive line ran "(plus (atomcount (car s)) (atomcount (cdr s)))". Here it will be quite analogous. We shall have a recur-

```
n=0:  tato

n=1:  tato
       (and tato only)

n=2:  tato
       (and tato only)
        (and tato (and tato only) only)

n=3:  tato
       (and tato only)
        (and tato (and tato only) only)
         (and tato (and tato only) (and tato (and tato only) only) only)
```

A short table of recursive expansions of the Lisp variable designated "tato" sive line featuring *two* calls on "replace", one involving the car of longlist and one involving the cdr, instead of just one involving the cdr. This makes perfect sense when you think about it. Suppose you wanted to replace all the unicorns in Europe with porpuquines. One way to achieve this nefarious goal would be to split Europe into two pieces: Portugal (Europe's car) and the remainder (Europe's cdr). After replacing all the unicorns in Portugal with porpuquines, and also all the unicorns in the rest of Europe with porpuquines, you would finally recombine the two new pieces into a reunified Europe. (This is supposed to suggest a "cons" operation.) Of course, to carry out this dastardly operation on Portugal an analogous splitting and rejoining would have to take place—and so on. That suggests our recursive line will look like this:

```
(cons (replace 'unicorn
               '(porpuquine)
               (car geographical-unit))
      (replace 'unicorn
               '(porpuquine)
               (cdr geographical-unit)))
```

Or, more elegantly and more generally:

```
(cons (replace atm lyst (car longlist))
      (replace atm lyst (cdr longlist)))
```

This "cons" line will cover the case where longlist's car is nonatomic as well as the case where it is atomic but not equal to atm. In order to make this work we need to augment the embryonic case slightly: we shall say that when longlist is not a list but an atom, "replace" has no effect on longlist at all. Conveniently this subsumes the earlier "null" line, and so we can drop that one. If we put all of this together, we come up with a new, improved definition:

```
(def replace
 (lambda (atm lyst longlist)
  (cond
   ((atom longlist) longlist)
   ((eq (car longlist) atm)
    (append
     lyst
     (replace atm lyst (cdr longlist))))
   (t
    (cons
     (replace atm lyst (car longlist))
     (replace atm lyst (cdr longlist)))))
 ))
```

Now when we say "(tato-expansion 2)" to the Lisp genie, it will print out for us the list "(tato (and tato only) (and tato (and tato only) only))".

Well, well. Isn't this a magnificent accomplishment? If it seems less than magnificent, perhaps we can carry it a step further. A recursive acronym—one containing a letter standing for the acronym itself—can be amusing, but what about *mutually* recursive acronyms? This could mean, for instance, two acronyms each of which contains a letter standing for the other acronym. An example would be the pair of acronyms "NOODLES" and "LINGUINI", standing respectively for:

"NOODLES (oodles of delicious
 LINGUINI), elegantly served"

"luscious itty-bitty NOODLES got
 usually in Naples, Italy"

Notice, incidentally, that "NOODLES" is not only indirectly recursive but also directly so. There is nothing wrong with that.

In general the notion of mutual recursion means a system of arbitrarily many interwoven structures each of which is defined in terms of one or more members of the system (possibly including itself). If we are speaking of a family of mutually recursive acronyms, this means a collection of words in any one of which letters can stand for any word in the family.

I have to admit that this specific notion of mutually recursive acronyms is not particularly useful in any practical sense. It is nonetheless quite useful as a droll example of a very common abstract phenomenon. Who has not at some time mused about the inevitable circularity of dictionary definitions? Anyone can see that all words eventually are defined in terms of some fundamental set that is not further reducible but simply goes round and round endlessly. You can amuse yourself by looking up the definition of a common word in a dictionary and replacing the main words in it by *their* definitions. I once carried this process out for "love," defined as "a strong affection for or attachment or devotion to a person or persons," substituting for "strong," "affection," "attachment," "devotion" and "person" and coming up with this concoction: "A morally powerful mental state or tendency, having strength of character or will for, or affectionate regard, or loyalty, faithfulness or deep affection to, a human being or beings, especially as distinguished from a thing or lower animal."

Not being satisfied with that, I carried the entire process a step further. This was my result: "A set of circumstances or attributes characterizing a person or thing at a given time in, with, or by the conscious or unconscious together as a unit full of or having a specific ability or capacity in a manner relating to, dealing with, or capable of making the distinction between right and wrong in conduct, or an inclination to move or act in a particular direction or way, having the state or quality of being strong in moral strength, self-discipline, or fortitude, or the act or process of volition for, or consideration, attention, or concern full of fond or tender feeling for, or the quality, state, or instance of being faithful to, those persons or ideals that one is under obligation to defend or support, or the condition, quality or state of being worthy of trust, or a strongly felt fond or tender feeling to a creature or creatures of or characteristic of a person or persons, that lives or exists, or is assumed to do so, particularly as separated or marked off by differences from that which is conceived, spoken of, or referred to as existing as an individual entity, or from any living organism inferior in rank, dignity, or authority, typically capable of moving about but not of making its own food by photosynthesis."

Isn't it romantic? It certainly makes "love" ever more mysterious. Stuart Chase, in his lucid classic on semantics, *The Tyranny of Words,* does a similar exercise for "mind" and shows its opacity equally well. Of course, concrete words as well as abstract ones get caught in this vortex of confusion. My favorite example is one I discovered while looking through a French dictionary many years ago. It defined the verb *clocher* ("to limp") as *marcher en boitant* (roughly "to walk while hobbling") and *boiter* ("to hobble") as *clocher en marchant* ("to limp while walking"). This eager learner of French was helped precious little by that particular pair of definitions.

But let us return to mutually recursive acronyms. I put quite a bit of effort into working out a family of them, and to my surprise they wound up dealing mostly (although by no means exclusively!) with Italian food. It all began when, inspired by "tato", I chose the similar word "tomato" and then decided to use its plural, coming up with this meaning for "tomatoes":

TOMATOES on MACARONI
 (and TOMATOES only),
 exquisitely SPICED

The capitalized words within the phrases are those that are also acronyms. Here is the rest of my mutually recursive family:

MACARONI:
 MACARONI and CHEESE
 (a REPAST of Naples, Italy)

REPAST:
 rather extraordinary PASTA
 and SAUCE, typical

CHEESE:
 cheddar, havarti, Emmentaler
 (particularly SHARP Emmentaler)

SHARP:
 strong, hearty and rather pungent

SPICED:

sweetly pickled in CHEESE
ENDIVE dressing

ENDIVE:
egg NOODLES, dipped
in vinegar eggnog

NOODLES:
NOODLES (oodles of delicious
LINGUINI), elegantly served

LINGUINI:
LAMB CHOPS
(including NOODLES),
got usually in northern Italy

PASTA:
PASTA and SAUCE (that's ALL!)

ALL!:
a luscious lunch

SAUCE:
SHAD and unusual COFFEE
(eccellente!)

SHAD:
SPAGHETTI, heated al dente

SPAGHETTI:
standard PASTA, always good, hot
particularly (twist, then ingest)

COFFEE:
choice of fine flavors,
particularly ESPRESSO

ESPRESSO:
excellent, strong, powerful, rich
ESPRESSO, suppressing sleep
outrageously

BASTA!:
belly all stuffed (tummy ache!)

LAMB CHOPS:
LASAGNE and meatballs,
casually heaped onto
PASTA SAUCE

LASAGNE:
LINGUINI and SAUCE
and GARLIC
(NOODLES everywhere!)

RHUBARB:
RAVIOLI, heated under butter
and RHUBARB (BASTA!)

RAVIOLI:
RIGATONI and vongole in oil,
lavishly introduced

RIGATONI:
rich Italian GNOCCHI
and TOMATOES
(or NOODLES instead)

GNOCCHI:
GARLIC NOODLES
over crisp CHEESE,
heated immediately

GARLIC:
green and red LASAGNE
in CHEESE

Any gourmet can see that little attempt has been made to have each term defined by its corresponding phrase; it is simply associated more or less arbitrarily with the phrase.

What happens if we begin to expand some word, say "pasta"? At first we get simply "PASTA and SAUCE (that's ALL!)". The next stage yields "PASTA and SAUCE (that's ALL!) and SHAD and unusual COFFEE (eccellente!) (that's a luscious lunch)". We could obviously go on expanding acronyms forever, or at least until we filled the universe to its very brim with mouth-watering descriptions of Italian food. But what if we were less ambitious and wanted merely to fill half a page or so with such a description? How might we find a way to halt this seemingly bottomless recursion in midcourse?

The key word here is "bottomless," and the answer it implies is: Put in a mechanism to allow the recursion to bottom out. The bottomlessness comes from the fact that at every stage every acronym is allowed to expand, that is, to spawn further acronyms. What if instead we kept tight control of the spawning process, being generous in the first few "generations" and gradually letting fewer and fewer acronyms spawn progeny as the generations got later? It would be similar to a redwood tree in a forest, which begins with a single "branch" (its trunk), and that branch spawns "progeny," namely the first generation of smaller branches, and they in turn spawn ever more progeny—but eventually a natural bottoming out comes as a consequence of the fact that teeny twigs simply cannot branch further. (Somehow in the course of evolution trees seem to have got their wires crossed, since for them bottoming out generally takes place at the top.)

If this process were completely regular, all redwood trees would look exactly alike and one could agree with former Governor Reagan's memorable dictum "If you've seen one redwood tree, you've seen them all." Unfortunately redwood trees (and maybe some other things) are trickier than Governor Reagan realized, and we have to learn to deal with a variety of things that go by the same name. The variety is caused by the introduction of randomness into the choices of whether to branch or not to branch, what angle to branch at, what size branch to grow and so on.

Similar remarks apply to the "trees" of mutually recursive acronyms. If in expanding "tomatoes" we always made exactly the same control decisions about which acronyms to expand when, there would be one and only one type of "rhubarb" expansion, so that here too it would make sense to say, "If you've seen one rhubarb, you've seen them all." But if we allow some randomness to enter the decision making about spawning, we can get many varieties of rhubarb, all bearing some telltale resemblance to one another but at a much more elusive level of perception.

How can we do it? The ideal concept to bring to bear here is that of the random-number generator, which serves as the computational equivalent of a coin flip or a dice throw. We shall let all decisions about whether or not to expand a given acronym depend on the outcome of such a virtual coin flip. At early stages of expansion we shall set things up so that the coin will be very likely to come up heads (do expand); at later stages it will be increasingly likely to come up tails (no expansion). The Lisp function "rand" will be employed for this. It takes no arguments, and each time it is called it returns a new real number somewhere between 0 and 1, unpredictably. (This is an exaggeration. It is actually 100 percent predictable if you know how it is computed, but since the algorithm is rather obscure, for most purposes and for most observers the behavior of the function will be so erratic that it counts as being totally random. The story of random-number generation is quite a fascinating one and would be an article in itself.)

If we want an event to happen with a probability of 60 percent, first we ask rand for a value. If the value turns out to be .6 or below, we go ahead, otherwise we do not. Since over a long period of time rand sprays its outputs uniformly over the interval between 0 and 1, we shall indeed get the go-ahead 60 percent of the time.

So much for random decisions. How do we get an acronym to expand when it is told to? That is not too hard. Suppose we let each acronym be a Lisp function, as in the following example:

```
(def tomatoes
 (lambda ()
  '(tomatoes on macaroni
   (and tomatoes only)
   exquisitely spiced))))
```

The function "tomatoes" takes no arguments and simply returns the list of words it expands into. Nothing could be simpler.

Now suppose we have a variable called "acronym" whose value is some particular acronym—but we do not know which one. How could we get that acronym to expand? The way we have set it up the acronym must act as a function call. In order for any atom to invoke a function it must be the car of a list, as in the examples "(plus 2 2)", "(rand)" and "(rhubarb)". If we were to write "(acronym)", the literal atom "acronym" would be taken by the Lisp genie

as a function name. That, however, would be a misunderstanding. It is certainly not the atom "acronym" that we want to make serve as a function name, but its *value,* be it "macaroni", "cheese" or what have you.

In order to accomplish this we resort to a little trick. If the value of the atom "acronym" is "rhubarb" and if I write "(list acronym)", the value the genie will return to me will be the list "(rhubarb)". The genie, however, will simply see this as an inert piece of Lispstuff rather than as a little command I should like to have executed. It cannot read my mind. How do I get it to perform the desired operation? The answer is that I remember the function called "eval", which makes the genie look on a given data structure as a wish to be executed. In this case I need merely say "(eval (list acronym))" and I shall get the list "(ravioli, heated under butter and rhubarb (basta!))". And if the acronym had had a different value, the genie would have handed me a different list.

We now have just about enough ideas to build a function capable of expanding mutually recursive acronyms into long but finite phrases whose sizes and structures are determined by many "flips" of the "rand" coin. Instead of stepping you through the construction of this function I shall simply display it and let you peruse it. It is modeled very closely on the earlier function "replace":

```
(def expand
 (lambda (phrase probability)
  (cond
   ((atom phrase) phrase)
   ((is-acronym (car phrase))
    (cond
     ((lessp (rand) probability)
      (append
       (expand (eval (list (car phrase)))
               (lower probability))
       (expand (cdr phrase) probability)))
     (t
      (cons (car phrase)
            (expand (cdr phrase)
                    probability)))))
   (t
    (cons
     (expand (car phrase)
             (lower probability))
     (expand (cdr phrase) probability))))
 ))
```

Note that "expand" has two parameters. One parameter represents the phrase to be expanded, the other represents the probability of expanding any acronyms that are top-level members of the given phrase. (Thus the value of the atom "probability" will always be a real number between 0 and 1.) As in the redwood-tree example, the expansion probability should decrease as the calls get increasingly recursive. This is why lines that call for the expansion of (car phrase) do so with a *lowered* probability. To be exact, we can define the function "lower" as follows:

```
(def lower (lambda (x) (times x .8)))
```

Thus each time an acronym expands, its progeny are only .8 times as likely to expand as it was. This means acronyms nested deep enough have a vanishingly small probability of spawning further progeny. You could use any reducing factor; there is nothing sacred about .8 except that it seems to yield good results for me.

The only remaining undescribed function inside the definition above is "is-acronym". Its name is pretty self-explanatory. First the function tests to see if its argument is an atom; if it is not, the function returns nil. If the argument *is* an atom, the function goes on to see whether that atom has a function definition, in particular a definition with the form of an acronym. If the atom has such a definition, is-acronym returns the value t; otherwise it returns nil. Precisely how this is done depends on your specific variety of Lisp, which is why I have not shown it explicitly. In Franz Lisp only one line is needed.

You may have noticed that there are two cond clauses in close proximity that begin with "t". How come one "otherwise" follows so closely on the heels of another one? Actually they belong to different conds, one nested inside the other. The first "t" (belonging to the inner cond) applies to a case where we know we are dealing with an acronym but where our random coin, instead of coming down heads, has come down tails (which amounts to a decision not to expand); the second "t" (belonging to the outer cond) applies to a case where we have discovered we are simply not dealing with an acronym at all.

The inner logic of "expand", when it is scrutinized carefully, makes perfect sense. On the other hand, no matter how carefully you scrutinize it, the output produced by "expand" using this "famiglia" of acronyms remains quite silly. Here is an example:

(rich italian green and red linguini and shad and unusual choice of fine flavors, particularly excellent, strong, powerful, rich espresso, suppressing sleep outrageously (eccellente!) and green and red lasagne in cheese (noodles everywhere!) in cheddar, havarti, Emmentaler (particularly sharp Emmentaler) noodles (oodles of delicious linguini), elegantly served (oodles of delicious linguini), elegantly served (oodles of delicious linguini and sauce and garlic (noodles (oodles of delicious linguini), elegantly served everywhere!) and meatballs, casually heaped onto pasta and sauce (that's all!) and sauce (that's a luscious lunch) sauce (including noodles (oodles of delicious linguini), elegantly served), got usually in northern Italy), elegantly served over crisp cheese, heated immediately and tomatoes on macaroni and cheese (a repast of Naples, Italy) (and tomatoes only), exquisitely sweetly pickled in cheese endive dressing (or noodles instead) and vongole in oil, lavishly introduced, heated under butter and rich Italian gnocchi and tomatoes (or noodles instead) and vongole in oil, lavishly introduced, heated under butter and rigatoni and vongole in oil, lavishly introduced, heated under butter and ravioli, heated under butter and rich Italian garlic noodles over crisp cheese, heated immediately and tomatoes (or noodles instead) and vongole in oil, lavishly introduced, heated under butter and ravioli, heated under butter and rhubarb (basta!) (basta!) (basta!) (basta!) (belly all stuffed (tummy ache!)) (basta!))

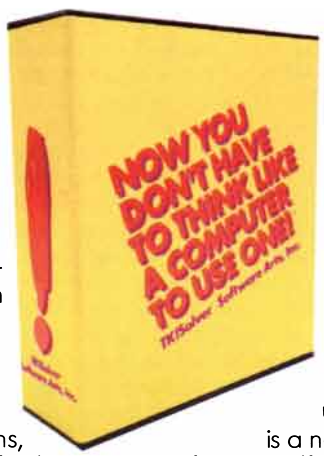
Can you figure out which acronym this gastronomical monstrosity grew out of? As you can see, Lisp is hardly the computer language to learn if you want to lose weight. One final example of the glories of recursive spaghetti is given herewith, expanded from a different starting point:

(macaroni and cheese (a rather extraordinary pasta and sauce, typical of Naples, Italy) and cheddar, havarti, Emmentaler (particularly sharp Emmentaler) (a rather extraordinary pasta and shad and unusual coffee (eccellente!) (that's a luscious lunch) and sauce (that's all!) and shad and unusual choice of fine flavors, particularly espresso (eccellente!) (that's all!) and sauce (that's a luscious lunch) and spaghetti, heated al dente and unusual choice of fine flavors, particularly excellent, strong, powerful, rich espresso, suppressing sleep outrageously (eccellente!), typical of Naples, Italy))

The "expand" function exploits one of the most powerful features of Lisp. That is the ability of a Lisp program to take data structures it has created and treat them as pieces of code (that is, give them to the Lisp genie as commands). Here it was done in a most rudimentary way. An atom was wrapped in parentheses and the resulting minuscule list was then evaluated, or "eval'd", as Lispers' jargon has it. The work involved in manufacturing the data structure was next to nothing in this instance, but in other instances elaborate pieces of structure can be "consed up", then handed to the Lisp genie for "eval'ing". Such pieces of code might be new function definitions or any number of other things. The main idea is that in Lisp one has the ability to "elevate" an inert, information-containing data structure to the level of an "animate agent," where it becomes a manipulator of inert structures itself. This program-data cycle, or loop, can con-

CRUNCH!

The inventor of VisiCalc,® the world's leading software product, introduces TK!Solver™—the revolutionary way to crunch problems on a personal computer.



With VisiCalc, Software Arts™ created an entirely new application for personal computers—the electronic spreadsheet.

Now Software Arts has developed TK!Solver, another entirely new and different application for personal computers—the problem solver. Business and financial problems. Scientific and engineering problems. Architectural and building design problems. Any kind of problem you can dream up, in any area.

What's so revolutionary about that? Simply this: the power of the TK!Solver program comes from the ease with which you can set up problems, vary assumptions, find solutions, and display results. All the facilities needed to solve the problems—simple or complex—are built-in, and need not be developed. Once you have stated your problem, simply enter the known values, then solve the problem with a single keystroke.

For even easier problem solving use a TK!SolverPack™ in which models for solving common problems in such areas as finance, engineering and education have already been formulated by professionals in these fields.

There is no software product in existence that solves problems as quickly, flexibly, and easily as TK!Solver—The Problem Cruncher™. In fact, no other software product even works the way TK!Solver does. It is a necessity for personal computer owners. See for yourself. Visit your nearest computer store for a demonstration and find out how much you can do with a personal computer when it thinks as flexibly as you do.

Software Arts™

27 Mica Lane, Wellesley, MA 02181

tinue on and on, with structures reaching out, twisting back and indirectly modifying themselves or structures related to them.

Certain types of inert, or passive, information-containing data structures are sometimes referred to as "declarative knowledge"—"declarative" because they often have a form abstractly resembling that of a declarative sentence, and "knowledge" because they encode facts about the world, accessible by looking in an index in somewhat the way "book-learned" facts are accessible to a human being. In contrast, animate, or active, pieces of code are referred to as "procedural knowledge"—"procedural" because they define sequences of actions ("procedures") that actually manipulate data structures, and "knowledge" because they can be viewed as embodying the program's set of skills, something like a human being's unconscious skills that were once learned through long rote drill sessions. Sometimes these contrasting knowledge types are referred to as "knowledge that" and "knowledge how."

This distinction should remind biologists of the distinction between genes, which are relatively inert structures inside the cell, and enzymes, which are anything but inert. Enzymes are the animate agents of the cell; they transform and manipulate all the inert structures in indescribably sophisticated ways. Moreover, Lisp's loop of program and data should remind biologists of the way genes dictate the form of enzymes and enzymes manipulate genes (among other things). Thus Lisp's procedural-declarative program-data loop provides a primitive but very useful and tangible example of one of the most fundamental patterns at the base of life: the ability of passive structures to control their own destiny, by creating and regulating active structures whose form they dictate.

We have been talking all along about the Lisp genie as a mysterious given agent, without asking where it is to be found or what makes it work. It turns out that one of Lisp's most exciting properties is that one can describe with ease the Lisp genie's complete nature in Lisp itself. In other words, the Lisp interpreter can be easily written down in Lisp. Of course, if there is no Lisp interpreter to run *that* Lisp interpreter, it might seem like an absurd and pointless exercise, a bit like having a description in flowery English telling foreigners how best to learn English. It is not, however, as silly as that makes it sound.

In the first place, if you know enough English, you can "bootstrap" your way further into English; there is a point beyond which explanations written in English about English are indeed quite useful. What is more, that point is not too far beyond the beginning level. Therefore all you need to acquire first, and autonomously, is a "kernel"; then you can begin to lift yourself by your own bootstraps. For children it is an exciting thing when, as they read, they begin to learn new phrases all by themselves, simply by encountering them several times in succession. Their vocabulary begins to grow by leaps and bounds. So it is once there is a Lisp kernel in a system; the rest of the Lisp interpreter can be written in Lisp and usually is.

The fact that one can easily write the Lisp interpreter in Lisp is no mere fluke of some peculiarly introverted fact about Lisp. The reason it is easy is that, because of its program-data loop, Lisp lends itself to writing interpreters for all kinds of computer languages. This means Lisp can serve as a basis on which one can build other languages.

To put it more vividly, suppose you have designed on paper a new language called "Flumsy." If you really know how Flumsy should work, it should not be too hard for you to write an interpreter for it in Lisp. Once your Flumsy interpreter is implemented, it becomes in essence a new genie to which you can give wishes in Flumsy and that will in turn communicate those wishes to the Lisp genie in Lisp. Of course, all the mechanisms allowing the Flumsy "meta-genie" to talk with the Lisp genie are themselves being carried out by the Lisp genie. Is this a mere facade? Is talking Flumsy really just a way of talking Lisp in disguise?

Well, when the U.S. arms negotiators talk with their Russian counterparts through an interpreter, are they really just speaking Russian in disguise? Or is the crux of the matter whether the interpreter's native language was English or Russian, on which the other language was built as a second one? And suppose you find out that actually the interpreter's native language was Lithuanian, that she learned English only as an adolescent and then learned Russian by taking high school classes taught in English? Will you then think that when she speaks Russian, she is actually speaking English in disguise, or worse, that she is actually speaking Lithuanian, doubly disguised?

Analogously, you might find that the Lisp interpreter is actually written in Pascal or some other computer language. And then someone could strip off the Pascal facade as well, revealing to you that all instructions are really being executed in *machine language*, so that you are fooling yourself completely if you think the machine is talking Flumsy, Lisp, Pascal or any other higher-level computer language!

When one interpreter runs on top of another, there is always the question of what level one chooses not to look below. I personally seldom think about what underlies the Lisp interpreter, so that when I am dealing with the Lisp

Phone Extortion

Let's put the screws to our phone suppliers on a terrific long range cordless phone with a novel new consumer buying tool called GREED!

It's a war. The major phone suppliers are each desperately trying to lock up big orders. And now, we've devised a plan to take advantage of their greed.

Together we can apply enough pressure to get you a super high quality cordless phone with a range up to 700' for just $79. And, hopefully we'll make a good profit for ourselves. But, there's a problem.

TIGHTEN THE SCREWS

Before I tell you all about the great features and sound of the phone, let me explain just how we plan to deliver these phones for such an admittedly foolhardy price. You see, it's called extortion (legal definition please) and it works like this.

There are two large phone companies that are trying to get DAK's phone business. To be perfectly honest, the quality, guarantee and sound are so close that we couldn't decide between the two.

Even their prices are close. So here's what we did. We got quotes from both of them for 5,000, then for 50,000 phones.

And, here's what we decided. As you place your orders, we'll offer each of the two companies a check (cash in advance) for the number of phones we need.

The company with the **lowest price of the day** will get that day's order. You see, your orders give us incredible leverage.

Usually we buy just one phone from one supplier. Unfortunately once they have a contract with us, they have no incentive to lower our price. And, at our costs today, we really can't come out at only $79.

SAFE INVESTMENT

Don't worry about your $79. DAK is a large company. If this plan to sell 50,000 phones fizzles out, we'll still be OK. Plus, DAK doesn't even charge your credit card until **after** we ship.

And most important. Each phone will come to you in its factory sealed carton, and will be backed by the manufacturer's standard limited warranty.

Finally, you won't be getting a cheap stripped down phone. We aren't dumb. We picked the phones we wanted **before** we got the quotes. So, nobody can make any changes to come out lowest.

And frankly we'll get 50,000 of you new customers (we have 225,000 now) to send our 68 page electronics catalog to. So, even if it does fizzle, we're sure to get another chance at you later.

A GREAT LONG RANGE PHONE

So, now let's see if this is really going to be a stripped down low quality phone. Remember, each and every feature I describe appears on both phones. And, if for some emotional reason you want to specify which phone you want, no problem, I've got a way for you to do that too.

Imagine walking from room to room in your home talking on the phone. You'll go out and get the mail, walk over to a neighbors or work in the yard or the garage.

It's a superb personal security device. You can instantly call a doctor or police from anywhere you happen to be.

And, if you are needed at home, no problem. There's a page button on the base that beeps you at the handset which then becomes a wireless intercom.

Cordless phones are a great combination of posh luxury and life saving utility.

NO INSTALLATION

If you can plug in a lamp you can install your new cordless phone. Just plug in the modular phone plug and the AC line cord, then start talking. The phone uses universal pulse dialing compatible with both rotary dial and Touch Tone® phone lines.

LOADED WITH FEATURES

Look at these features. You'll have last number redial to redial busy numbers. And mute lets you electronically put your hand over the mouthpiece for privacy.

There's both a standby/talk switch and a power off switch so that you don't have to return the handset to the base when you aren't using it.

And when you do put the handset in the base, its heavy duty NiCad batteries recharge automatically, and the auto secure protects your base station.

If you're adjusting a TV antenna, the person at the base can talk to you using the intercom feature and any of your regular house phones. So, you'll have a cordless intercom, with a range up to 700 ft too.

The two companies that we are trying to 'extort' our low prices from are Unitech, and U.S. Tron. Even their names sound alike. Both phones are backed by a manufacturer's standard limited warranty.

TRY TOTAL PHONE FREEDOM RISK FREE

Experience the thrill of total phone freedom as you roam throughout your home, yard or even a neighbor's house. You'll never have to 'run for the phone' again.

Try a cordless phone in your own home risk free. If you don't find the luxury of talking from wherever you are 100% enthralling, just return the phone to DAK in its original box within 30 days for a refund.

To order a long range full featured cordless phone risk free with your credit card, call toll free or send your check for the 'extorted' price of just **$79** plus $3.50 for postage and handling. Order Number 9615. CA res add 6% sales tax.

If you have a particular love for either phone, you can have your choice for just $89 plus $3.50 for postage and handling. To order the U.S. Tron, use Order Number 9616. To order the Unitech, use Order Number 9617. So, you'll pay only $79 if the choice is ours, or $89 if it's yours.

DAK
INDUSTRIES INCORPORATED
TOLL-FREE 1-800-423-2636
If busy, after hours, on weekends or in CA
CALL TOLL-FREE . . . 1-800-228-1234
10845 Vanowen St., N. Hollywood CA 91605

system, I feel as if I am talking with "someone" whose "native language" is Lisp. Similarly, when I am dealing with people, I seldom think about what their brains are composed of; I do not reduce them mentally to piles of patterned neuron firings. It is natural to my perceptual system to recognize them at a certain level and not to look below that level.

If someone were to write a program that could deal in Chinese with simple questions and answers about restaurant visits, and if that program were written in the hypothetical language "SEARLE" (for "Simulated East-Asian Restaurant-Lingo Expert"), I could choose to view the system either as genuinely speaking Chinese (assuming it gave a creditable and not too slow performance) or as genuinely speaking SEARLE. I can shift my point of view at will. The one I adopt is governed mostly by pragmatic factors, such as which subject area I am currently more interested in (Chinese restaurants or how grammars work), how closely matched the given level's speed is to that of my own brain and, not least, whether I happen to be more fluent in Chinese or in SEARLE. If to me Chinese is a mere bunch of squiggles and squoggles, I may opt for the SEARLE viewpoint; if on the other hand SEARLE is a mere bunch of confusing technical expressions, I shall probably opt for the Chinese viewpoint. And if I find out that the SEARLE system is in turn implemented on top of a Lisp system, then I have yet a third point of view to choose. And so on.

With interpreters stacked on interpreters, however, things rapidly become very inefficient. It is like running a motor on power coming from a series of electric generators, each generator being run on power coming from the preceding one: a good deal of power is lost at each stage. With generators there is usually no need for a long cascade, but with interpreters it is often the only way to go. If there is no machine whose machine language is Lisp, then you build a Lisp interpreter for whatever machine you have available and run Lisp that way. And Flumsy or SEARLE, if you want to have it at your disposal, is then built on top of this "virtual Lisp machine." Such indirectness can be annoyingly inefficient, causing your new virtual Flumsy machine or virtual SEARLE machine to run dozens of times slower than you would like.

There have been important hardware developments in the past several years, and now machines are available that are based on Lisp at the hardware level. This means that they "speak" Lisp in a somewhat deeper sense—let us say "more fluently"—than virtual Lisp machines. It also means that when you are working on such a machine, you are "swimming" in a Lisp environment. A Lisp environment goes considerably beyond what I have described so far, because it is more than just a language for writing programs. It includes an editing program, with which one can create and modify one's programs (and text as well), a debugging program, with which one can easily localize one's errors and correct them, and many other features, all designed to be compatible with one another and with an overarching "Lisp philosophy."

Such machines, although they are still expensive and somewhat developmental, are rapidly becoming cheaper and more available. They are put out by various new companies such as LMI (LISP Machine, Inc.), Symbolics, Inc., both of Cambridge, Mass., and older companies such as Xerox. Lisp is also available on most personal computers; you need only look at any issue of the current microcomputer magazines to find advertisements for it.

Why, in conclusion, is Lisp popular in artificial intelligence? There is no single answer, or even a simple answer. Would it be valid to say Lisp is "the language of thought"? Certainly not. AI people might have said so 10 or 20 years ago, but few would be so bold today. One reason Lisp has remained so popular is, as I wrote in my first Lisp column, that Lisp is crisp. Indeed, Lisp can be so elegant that a good Lisp function can please a Lisper as a poem pleases a lover of poetry. My colleague Dan Friedman is in fact compiling a book titled *Lisp Poems,* and he urges interested readers to send him their most elegant efforts. His address is Computer Science Department, Indiana University, Bloomington, Ind. 47405.

What properties of Lisp, then, have made it central to AI research? I would say the following ones are nearly exhaustive:

1. Lisp is crisp.
2. Lisp is interactive.
3. Lisp is centered on the idea of lists and their manipulation, and lists are extremely flexible, fluid data structures.
4. Lisp code, having the same form as Lisp data, can easily be manufactured in Lisp and run.
5. Interpreters for new languages can easily be built and experimented with in Lisp.
6. "Swimming" in a Lisplike environment feels natural for many people.
7. Lisp is permeated by the "recursive spirit."

Perhaps it is this last rather intangible statement that gets at it the best. For some reason many people in artificial intelligence seem to have a deep sense that recursivity, in *some* form or other, is connected with the "trick" of intelligence. It is a hunch, an intuition, a vaguely felt and slightly mystical belief, and one that I certainly share, but whether it will pay off in the long run remains to be seen.

BOOKS

Crystals at high pressure, salt hunger, our star the sun, blueprints, the piano

by Philip Morrison

COMPARATIVE CRYSTAL CHEMISTRY: TEMPERATURE, PRESSURE, COMPOSITION AND THE VARIATION OF CRYSTAL STRUCTURE, by Robert M. Hazen and Larry W. Finger. John Wiley & Sons, Inc. ($43.95). Their eyes on the solid earth, mineralogists must pursue the behavior of mineral crystals under high pressure. A hundred kilobars (one bar is an atmosphere of load) takes the sample to an equivalent depth of 300 kilometers, within the earth's mantle, far beyond the dreams of the drillers. In this remarkable summary of the work of a decade or so, two deft and ingenious experimenters at the Carnegie Institution of Washington let us into a body of techniques all but magical in their simplicity and power. The second half of their book is a wide synopsis of the data obtained. The general reader will perhaps enjoy most the step-by-step explications of how it is all done.

The work of the high-pressure pioneers, like Percy Bridgman between the wars, was clever, heroic, clumsy. The big hydraulic press, kept behind a thick wall in case of catastrophic failure, strained against a complex fitting deep within which some sample was under stress. The behavior of the sample was monitored by a wild variety of obscure effects that could somehow be teased to signal out of the hidden region along a wire. Those measurements did not often reach the 100-kilobar range. Nowadays the work appears effortless, carried out under these crushing pressures in palm-sized apparatus loaded by tightening up by hand with a Crescent wrench!

The preconditions for this dazzling development are clear. The fundamental physical understanding is that a single crystal a tenth the bulk of a pinhead has enough atoms to display with precision the X-ray-diffraction pattern that reveals its atomic structure and spacing. The chief technical foundation is the long development of X-ray diffraction, so that nowadays automatic, sensitive photon-counting instruments provide speedy printout of structural information at high precision in hours, where older methods would have taken months of hand setting and hand computation. Finally, the strong bonds of the lightest atoms, realized in diamond and in boron carbide, supply us with materials that are the hardest, yet are entirely transparent to the X rays used to probe most important minerals.

To make a high-pressure, high-temperature cell able to hold a crystal in the analyzing beam it is enough to begin with a pair of quarter-inch slices cut out of hardened steel rod stock an inch and a half in diameter. Three half-inch cap screws serve to load the sandwich. Fitted with care into the steel disks is a multiple filling, first two boron carbide disks, then a refractory ring with heater wires. The tiny sample, along with two or three calibrating crystals, lies in the pressurized core. That cell is a one-third-millimeter hole in a foil of nickel alloy of about the same thickness. The ceiling and floor of the cell are two carefully prepared diamond anvils, borne between the boron carbide disks.

A methanol-based fluid fills the cell. Turn the cap screws in sequence a few degrees at a time, switch on the heater control circuitry and you can place in the path of a fine diffractometer an X-ray-transparent sample at the conditions of the earth's mantle. The diamond anvils, the heart of the device, must be well made, but they are not large enough for a modest engagement ring. The force they transmit is a few hundred pounds; the stored energy released when they fail is no larger than that of a laboratory notebook falling to the floor. The heroics have been supplanted by subtlety.

The technique is not exhausted. Hazen and Finger, the authors of this book, are working with crystallized gases, for example argon and neon, to replace the methanol pressure fluid. That change promises a pressure range a few times higher, at least at cryogenic temperatures. A tiny diamond-anvil system of somewhat more complicated mechanical design has reached the highest static pressures ever attained: close to two megabars.

The authors have made their book a how-to-do-it one. It has the flavor of old days at the laboratory bench, with its careful reminders of the right use for Vaseline, Kleenex, acetone, wax and cements, and close microscopic inspection as the marvelous structure is being made up out of its apt and precious parts. The second half of the book discusses results. A few dozen researches, often on a related family of inorganic compounds, are reported in summary, with tables and graphs. The authors have included all the high-temperature, high-pressure X-ray structure data they knew of up to mid-1981.

The results are satisfying: the unit polyhedrons that form many of the important oxides and silicates in this study allow rough prediction of the variations of structures with ambient conditions and with systematic chemical composition change. That justifies the title of the volume. Phase changes, complex and important as they are, are not yet so easy to tame, but continuous changes in structure suggest the rise of misfit between neighboring geometrical elements as a harbinger of phase change.

It is hard to name another domain of physics in which so much power has flowed from such simplicity and propriety of means. The topic, like its tools, is a small diamond of the laboratory.

THE HUNGER FOR SALT: AN ANTHROPOLOGICAL, PHYSIOLOGICAL AND MEDICAL ANALYSIS, by Derek Denton. Springer-Verlag ($149.80). Going beyond a monograph, here is a treatise. Professor Denton is a distinguished experimental endocrinologist, director of the Florey Institute in Melbourne. There he has worked for 30 years with the docile sheep, surgically replumbed to offer answers to the interplay of need and supply, cue and behavior, in the intake and outflow of sodium ion. On this firm foundation of artful experiment and dauntless technique he has built a book with 10,000 citations, since he is as much at home in the library as he is in the surgery and the metabolism cages.

The topic is recognizably wide. The initial lines of text recall how Humphrey Davy "bounded about the room in ecstatic delight" once his great battery had isolated the first tiny globules of the two main alkaline metals. Sodium is the "main cation of body fluids," as salt flavors all the world's ocean. Inland of the salt-charged sea breeze, land life must somehow find sodium where only traces exist. How that happens, particularly with our own species, with the physiological background resting on the work with sheep and rats, and on clinical studies in the hospital and afield, fills 650 pages of a readable, reflective and marvelously catholic work.

Consider the relevance, and yet the unexpectedness, of documented and appraising accounts of cannibalism, of pica, an appetite for strange substances, of the cafeteria experiments, with sheep or children free to choose an offered variety of foodstuffs, of the salty diet of farming villages of Honshu, full of soy

W I N E & F O O D

Cabernet Sauvignon

Cabernet Sauvignon is the unrivaled King of California wine grapes. The distinction is well deserved for it was this magnificent varietal that first put California on the world map of wine. A native of Bordeaux, the wines produced of this grape have, for centuries, been among the world's finest and most enduring.

WINE

As with all heartier red wines, Cabernet Sauvignon must be properly aged to be at its best. When fully mature, a deep aromatic bouquet typically is followed by round, full flavors suggesting berries and herbs, warm earthy flavors, rich in varietal character. Velvety smooth in texture, it is a gentle giant of a wine.

Recognized for its excellence, our gold medal winning 1977 Proprietor's Reserve Cabernet Sauvignon

is classic in its composition and sufficiently aged to be entirely enjoyable now.

FOOD

Traditionally, a wine as full bodied and flavorful as Cabernet Sauvignon is served with dishes of beef or venison, hearty stews, cheeses and rich pastas.

In our home, we enjoy Cabernet with a wide variety of foods. One I'm especially fond of is our own *Eagle Ranch Roast*, a wine country dish developed by my wife Vicki. Generously seasoned with olive oil, garlic and rosemary, this complex layering of flavors would overpower a wine of lesser body and character.

If you would enjoy sampling some of Vicki's recipes and learning more about wine, please write to the address below for your free brochure, "Wine and Food: Cabernet Sauvignon" a collection of family favorites uniquely compatible with this King of California Wines.

Discover Sebastiani Vineyards, a family winery in the Sonoma Valley.

Sam J. Sebastiani

Sam J. Sebastiani

Sebastiani Vineyards, Dept. G, P.O. Box AA, Sonoma, CA 95476 sauce, pickles and miso, and of the role of fire in the entry of hominids into the Temperate Zone. It was not warmth that counted but glow, with its ability to break the rigid circadian rhythm that then required, it is held, a strict 12-hour duration of sleep throughout the tropical night.

Or look at the pictures. They include a wild kangaroo at twilight in the Snowy Mountains, visiting the filter-paper cafeteria blocks that offered a choice of water or four distinct salts. (Sodium was the runaway favorite.) Five thousand camels are shown being assembled in Algeria for the 600-mile march south across the central Sahara to fetch salt from the ancient mines of Mali. The modified sheep of the Melbourne laboratory are here, their saliva diverted, an adrenal or a kidney carried now in a fold of neck skin convenient for valving and sampling. Phosphate is rarer than sodium; carnivores, Lapland reindeer and the cattle of the King Ranch alike seek it by long chewing on the bones they find on the deficient pasture. A series of autoradiographs that record the sites of specific protein synthesis is presented to illustrate the future direction of endocrine studies.

The still uncertain implications for medicine are worth summary; that they are tentatively and prudently advanced must go without saying. The measured output of essential sodium, like the input, varies among peoples from the extreme salt-free life of the fierce Yanomamö of the Amazon to the salt-doting Japanese towns of the north by a neat factor of 100. Some effects appear at these extremes, although they are only marginally clear. Over a wide range of intake, however, the sodium content of the human body remains in balance, stabilized by diverse feedback loops. The hormonal control is clear, if in detail we understand it only in part.

Animals are led to take required salt by instinct, responsive to its taste and not to a learned sense of well-being. A conjecture suggests that the signal of balance is the salt content not of the body as a whole but of specialized neurons that somewhere determine the taste for salt. Above this physiological level the learned response that most complex societies have nurtured encourages salt consumption beyond need by up to two orders of magnitude. The system adjusts, but at a cost of hormone manufacture up by a large factor, and by other consequences we may not yet perceive. One suggestion is strong: those Japanese villagers have an endemic high blood pressure. They die by cerebrovascular accident at a much higher rate than other populations. The matter is not resolved; epidemiology still offers "the Scottish court verdict: not proven." Nevertheless, a preventive approach might be a very "good investment" in a situation where a fifth of the people in the developed countries face circulatory-system hazard.

THE SUN, OUR STAR, by Robert W. Noyes. Harvard University Press ($20). If ever we are to understand our past and our future, we must learn to look at the ancient sun in new ways. That is the lesson a reader takes first of all from this well-written account of solar physics and astronomy today, done in satisfying depth without one equation, although with plenty of graphs and diagrams.

The new ways enrich this tale, which by no means neglects the basics of spectral lines and atomic abundances, of eclipse tracks and sunspot cycles. It is something of a surprise to see a solar telescope on the icy plateau at the South Pole itself. The location is natural enough: there in midsummer the sun never sets. Through five 24-hour days an international team watched the sun uninterruptedly. (In Christmas week the weather at the pole is favorable: the temperature rises often to a summery 30 degrees below zero F.) The aim is to record in detail the velocity patterns of the solar gases averaged over a large area of the sun's surface. Single-wave trains can be followed there for days on end. The results amount to solar seismology, the study of oscillations that extend deep into the sun.

For a long time we have recorded the "solar constant"—the total radiation arriving at the earth from the sun—from the thin air of the best mountain stations. That air is clear but not clear enough; in 1980 the solar input was measured from orbit, beyond the whole of the atmosphere, with an unprecedented accuracy of one part in 10,000. The sun was much mottled, and the solar constant went down markedly once a large sunspot group rolled into view. Before that natural-enough result we could not know whether some slight general brightening might not compensate for the obvious dark spots.

Magnetic fields on the sun have been known almost from the beginning of the century. Sunspots, of course, are magnetic. Today we detect the looping and bridging fields on the solar surface at every scale: the high prominences, the pinpoint X-ray spots, the vast coronal holes out of which plasma currents stream into space, to generate magnetic storms and the subtle aurora once they engage the earth's own magnetic regime. The flares, sudden intensely radiating patches, are magnetic lightning, so to speak, on a solar scale. Although we do not yet understand the mechanism, the energy spent by flares must arise from a magnetic store established by the drifts and twists of the conducting solar plasma, as on the earth our Lilliputian flashes arise from an electrostatic store set up by insulating winds laden with charged droplets.

Inside the sun we cannot see. His birth and death are treated as stellar evolution foresees, supported more by the regularities in distant stars than by any special view of the sun. Here there is little news, and it is bad. The neutrino observatory, that tank watched so well by the radiochemists deep in the South Dakota lode, is our one channel into the solar depths, the only way we know to exploit the closeness of our star for information from its central nuclear furnaces. The result, however, does not fulfill the expectations of the theorists: the signal is small. Professor Noyes prudently sees the question as being unresolved; there is a growing feeling that the mystery is in fact an illusion. (Recent replication of the difficult nuclear experiments on which the input data for the calculations are based does not confirm the older results with the expected precision.) One inference from our privileged view of the sun has been carried out to the remote stars: a dozen years of careful monitoring of characteristic strong coronal emission lines has shown unmistakable starspot cycles, with the more mature stars displaying regular activity like the sun's.

A chapter is devoted to solar physics at the receiving end, that is, to the effects of solar energy on the earth. The suspicions that solar variations affect the earth's climate are well drawn, with the conclusion that while we do not understand how it can be, the evidence grows "that Nature is able to do what scientists cannot yet explain." It may be that this problem deserves priority beyond any other in astronomy and meteorology.

Text and illustrations join to set out a topic whose importance is as evident as its growth—and its incompleteness. Only one frustration mars a reader's pleasure in this straightforward and authoritative updating: scarcely any references are given (there is just one, in the preface), not even to the sources of the many figures and photographs, which much injures the welcome impression of timeliness and completeness. "Courtesy of..." phrases are simply not enough: we deserve to know who, how, when.

BLUEPRINTS: TWENTY-SIX EXTRAORDINARY STRUCTURES, by Christopher Gray, Patricia Brown and John Boswell. Simon and Schuster ($19.95). The Wright Flyer and the VW Beetle, the Hoover Dam and the Pentagon, the elephant-shaped house at Margate, N.J., and the Biltmore, a Loire chateau raised near Asheville, N.C., for a gilded Vanderbilt exemplify the extraordinary structures of this large paperback. Blueprints are reproductions of line drawings presenting the structure in simple projection. Characteristically they are works of the disciplined imagination:

"Microsoft® Multiplan™ allows me to explore more alternatives in less time. I think it leads to better management decisions."

Edward R. Schwinn, Jr.
President, Schwinn Bicycle Company

"Multiplan software helps me make better business decisions. It's as simple as that. With Multiplan on my microcomputer, I can explore a number of alternatives faster. That's really important to me in planning and tracking my totally handcrafted Schwinn Paramount Bicycle. As with anything handmade, cost control and product planning are major considerations. For instance, Multiplan lets me study the effect of the learning curve on overall productivity. And although we have a mainframe computer, I find that Multiplan is a tool that allows me to personally examine my business options—without spinning my wheels."

A tool for decision makers. Microsoft Multiplan is a microcomputer software program that can help you too. In many ways. You can analyze cash flow. Plan budgets. Forecast income. Manage production. Multiplan allows you to set up an electronic worksheet for whatever your need may be. And, it lets you explore alternatives faster because when you make one change you immediately see the results of that change throughout the worksheet and on any related worksheets. Just change one number and every number that depends on it is adjusted automatically.

In plain English. You won't have to learn a cryptic language to use Multiplan. It takes commands in plain English. And, it will prompt you as you go along by telling you what to do next. There's even a HELP key in case you need help at any point. Press it and it gives you information to get you going again.

Backed by our knowhow. There are a lot of good reasons to use Microsoft software. And among the best is "knowhow." Microsoft put BASIC, the first microcomputer language, on the very first microcomputer, and today, Microsoft software is running on well over a million computers. Worldwide.

See it in action. If you'd like to explore more alternatives yourself, ask your local computer dealer for a demonstration of Microsoft Multiplan. Multiplan is available for most popular personal computers, including Apple,® the IBM® PC and computers that use CP/M-80® or MS™-DOS operating systems. It's a good management tool for building bicycles. It's also a good management tool for building businesses. Like yours.

BETTER TOOLS FOR MICROCOMPUTERS

MICROSOFT™

MICROSOFT CORPORATION, 10700 NORTHUP WAY, BELLEVUE, WASHINGTON 98004

Microsoft is a registered trademark, and Multiplan, MS, and the Microsoft logo are trademarks of Microsoft Corporation. Apple is a registered trademark of Apple Computer, Inc. IBM is a registered trademark of International Business Machines, Corp. CP/M-80 is a registered trademark of Digital Research, Inc.

they are made in advance, tools for the task in hand, "links between idea and reality." As engineering gets more complex, blueprints grow voluminously, but interest, even as documentation, inheres in very few. The details of fastenings and piping are useful mainly to those who maintain and repair. Only the most general of views become documents of value, and even they are often lost in the disposal of unwanted bulk.

This delightful compilation has searched out blueprints in such variety that some are sure to appeal to any reader caught by the technical. The large contrasty pages of white-lined images on a deep blue ground are graphically attractive, their meaning quite apart. Each example is here given less abstract representation by a couple of excellent large photographs to display the completed project, so that the book is something of a self-teaching introduction to technical drawing as well.

The first entry is winning. It is the Statue of Liberty; we see a view from the harbor, an aerial view of the island with its star fort, and two blueprint pages of Mlle Liberty. They are revealingly intimate, with sections at each landing, and above all an elevation that displays her iron skeleton, so unexpected and so necessary a gift from M. Eiffel himself. These blueprints are not original construction drawings; they were made in 1942 as insurance against bomb damage, since the old plans were then beyond reach in occupied France.

Quite a few other examples share this slight loss of authenticity, having been prepared after the fact. The contoured site plan of the Hoover Dam, the tower of the Empire State rising high and thin across four big pages and the elegance of a hyperbolic cooling tower are striking as well. The cooling tower is accompanied by views of Three Mile Island. It may take a generation for our public to look at those beautiful forms again without a sense of foreboding, "majestic structures... tarred with a brush they do not deserve."

Even the big Blueprint Company of New York no longer makes blueprints proper. Since the 1930's drawings have been reproduced mainly by the dry diazo process and others, black line in one step and no shrinkage. The word "blueprint" has become a metaphor for any detailed plan and will long outlive the simple and beautiful blue artifacts first made by Sir William Herschel in about 1840. The compilers of this visual enchantment must be confirmed New Yorkers; they have located the Pentagon with the accuracy of one who would remark that Coney Island is along the New Jersey shore.

GIRAFFES, BLACK DRAGONS, AND OTHER PIANOS: A TECHNOLOGICAL HISTORY FROM CRISTOFORI TO THE MODERN CONCERT GRAND, by Edwin M. Good. Stanford University Press ($29.50). The conveyor belt bears baby grands into the distance. We see the assembly line of the Yamaha plant at Hamamatsu, today the largest producer of pianos in the world, however unfamiliar to Chopin and Fats Waller. Japan now makes about 40 percent of the nearly one million pianos made each year around the world, about what the U.S. and the U.S.S.R. manage together. Integrated, automated without the neglect of craft skills where they are indispensable, the Yamaha firm (Nippon Gakki Co., Ltd.) modestly sees itself more as the Toyota than as the Rolls of the modern pianoforte. It ought to come as no surprise that the well-made iron frames of Yamaha pianos are cast by the vacuum-shield mold process, first developed to improve detail in the admirable motorcycle engines made by the same company. It is just a century since pianos came to Meiji Japan.

This lighthearted but careful book tells how over three centuries the modern piano came to be, more as an example of the development of the piano as a machine than as part of social history or the history of music. Design, materials, manufacture all play distinct roles. The sources are plentiful, including a large scholarly literature, an anecdotal richness, the records of companies and tradesmen, and above all the archaeology of the pianos themselves in collections worldwide. Pianos are not sealed sites, in the archaeologist's jargon, and the instrument preserved as Beethoven's own might have been repaired and modernized since the composer's time.

Bartolomeo Cristofori served a musical Medici prince at Florence as harpsichord maker, a man of recognized talent. Although the big harpsichord could hold its own in volume with the ensembles of the day, instrumental or voice, it lacked dynamics. Only by choosing an alternate keyboard could the player vary either loudness or timbre. The prince himself set the craftsman to improving the harpsichord. In 1700 Cristofori completed a "new invention that produces *piano* and *forte*." The instrument was a harpsichord save in one vital respect: the new action, the mechanism that actuated the strings, flung a hammer up at the string with a speed smoothly depending on the force of the key stroke. The device could be heard, unlike the little tinkly clavichord, and it offered the dynamics the big twanging harpsichord lacked, a new counterpart to voice and bowed strings.

In his time Mozart composed for the classical piano; in 1777 he praised at length the instruments made by Stein at Augsburg: "The tone is always even. [My last sonata] sounds exquisite on Stein's pianoforte." By the close of the 18th century the classical piano had decisively displaced the harpsichord. Wide pitch range and big volume were newly desirable. Beethoven and his times—concert tours, a new public, a new forcefulness in music—were reflected in, although not directly responsible for, the new pianofortes. The finest London maker, Broadwood, sent the deaf Beethoven a new grand with a big tone. It came too late to be heard.

Pitch range and volume grew. Long and heavy strings meant much stored energy. Strong players such as Beethoven and Liszt broke strings in bunches. The tension began to warp the wood frames, braced as they might be. Between 1820 and 1840 or so iron entered and dominated the piano. Cast-iron frames held the tens of tons of total stress required by more than eight octaves of loud strings. Steel wire, annealed and drawn through dies of diamond and ruby, became available sometime before 1840; piano wire remains to this day an engineering standard of steel uniformity and strength.

The modern grand piano appeared in about 1860 under the name of Steinway & Sons, American innovators and world leaders for generations. The upright pianos are more diverse, if commoner, and they settled down somewhat later. Grotesque vertical forms like the giraffe were long gone. (In 1904 the American manufacturers collected and made a pyre of all the big black square pianos they could find.) Piano evolution had reached a plateau.

New urea-based glues, Teflon bushings at many bearing points and plastic keys mark the main novelties over the past century. The player piano won a brief ascendancy in the 1920's, to be all but expunged by the phonograph and the radio. Electronic pianos, both those that seek to resemble pianos in tone and those that do not, have budded but not yet flowered. The concert grand and the home piano differ now more than ever; so does their music.

The volume surveys its unexpected landscape without pretense and with careful attention to evidence. The place of acoustical science seems very small; the great Helmholtz had as students both Theodore Steinway and the inventor of an ingenious but never-popular new keyboard. Little seems to have come to this story, however, out of the laboratory, apart from materials. Indeed, a bit more physics would help the tale along. The notes and apparatus of the book are in general good, although one important page has been reduced to the ridiculous by an editorial oversight. An account of the piano action, not at all a simple mechanism, refers to part after part by the letters, from key *A* to jack spring *P*, but not one letter appears on the facing diagram. The editor responsible should be set to sanding a bushel of capstans, knuckles and wippens, whatever they are.

Take the 6-volume Microcomputing Library for only $2.95

when you join the Small Computer Book Club.

You simply agree to buy 3 more books—at handsome discounts—within the next 12 months.

This set of six carefully selected books—a $76.25 value—will be your basic reference source for small computing.

BASIC AND THE PERSONAL COMPUTER.
Thomas A. Dwyer and Margot Critchfield. Clear, down-to-earth introduction to programming in BASIC. Detailed examples and numerous illustrations explain the wide range of applications possible with personal computers. Also, many step-by-step examples of word processing, computer games, and computer simulations. **Publisher's Price $14.95**

USING CP/M: A Self-Teaching Guide.
J.N. Fernandez and R. Ashley. A complete introduction to the CP/M ("Control Program") software package used on many advanced microcomputers. How to use CP/M to copy files, edit, test programs, translate programs, and more—regardless of the programming language the microcomputer uses. **Publisher's Price $14.95**

MICROCOMPUTER GRAPHICS.
Roy E. Myers. Provides the essential mathematics and programming techniques you will need for computer graphics applications in BASIC, and shows you how to handle animated figures, peripheral equipment, colors and resolution, and many other topics. **Publisher's Price $11.95**

INTERFACING MICROCOMPUTERS TO THE REAL WORLD.
Murray Sargent III and Richard L. Shoemaker. The complete guide for computerizing your home, office, or laboratory. Tables, charts, and displays show you how to handle microcomputer software and operating systems; and how to monitor everything from the house lights to motors, switches, and display panels. Examples feature the Z-80 microprocessor. **Publisher's Price $14.50**

THE WORD PROCESSING BOOK: A Short Course in Computer Literacy.
Peter A. McWilliams. You'll see how to save up to $15,000 by buying a small computer with word processing capacities. Includes a brand name buying guide. **Publisher's Price $9.95**

EXECUTIVE COMPUTING.
John Nevelson. 25 business methods illustrated by computer programs, in a step-by-step casebook. You'll find details on pricing techniques...long-range planning...corporate data bases...and much more. **Publisher's Price $9.95**

SMALL COMPUTER BOOK CLUB is a division of the Library of Computer and Information Sciences, the oldest, largest and most respected computer book club in the world. The SMALL COMPUTER BOOK CLUB will keep you up-to-date with the latest developments in software, improvements in hardware, programming languages, and much more—all at handsome discounts.

So start enjoying the club's benefits today.

4 Good Reasons to Join

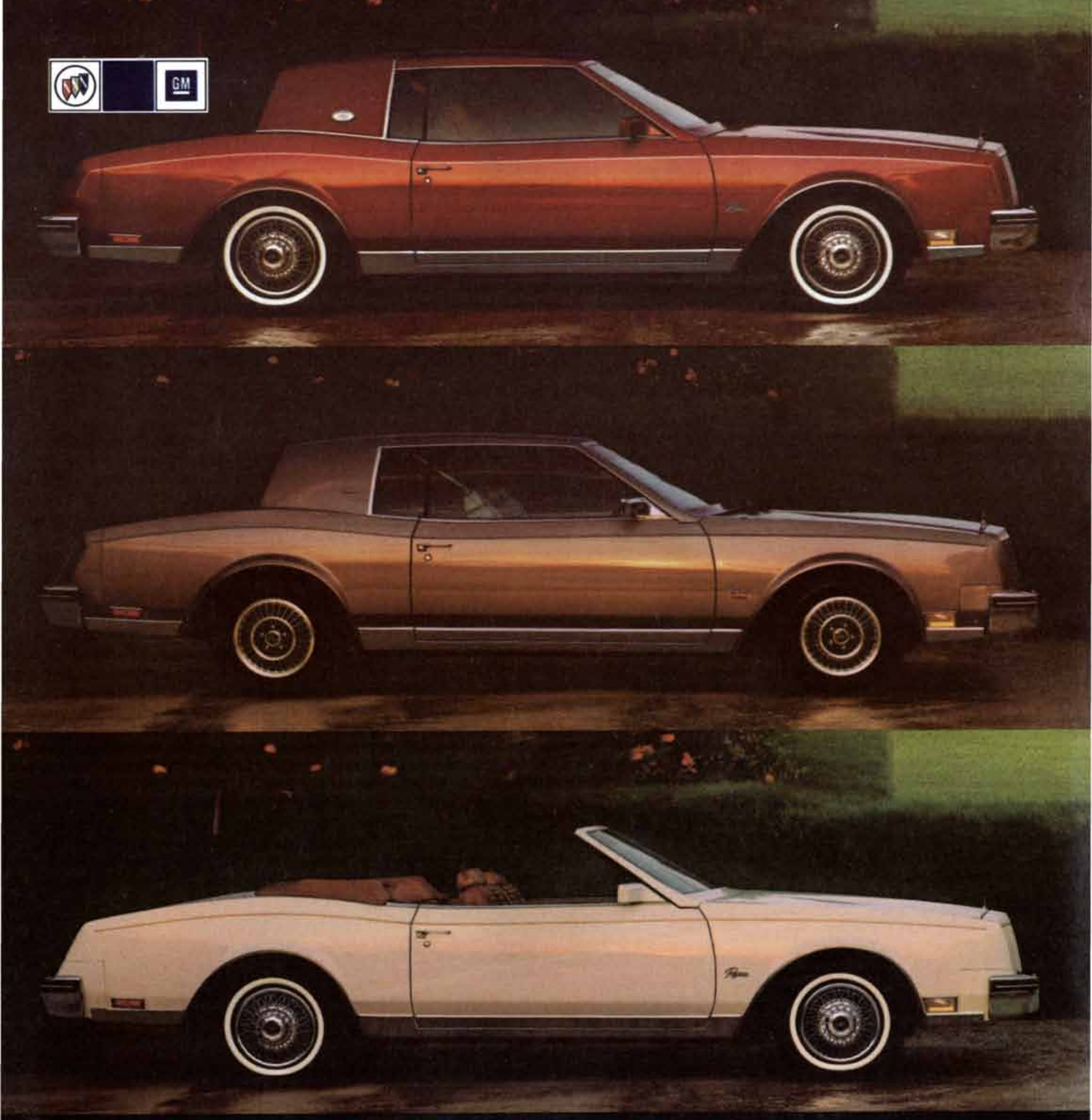
1. The Finest Books. Of the hundreds of books submitted to us each year, only the very finest are selected and offered. Moreover, our books are always of equal quality to publishers' editions, *never* economy editions.

2. Big Savings. In addition to getting the Microcomputing Library for $2.95 when you join, you keep saving substantially—up to 30% and occasionally even more. (For example, your total savings as a trial member—including this introductory offer—can easily be over 50%. That's like getting every other book free!)

3. Bonus Books. Also, you will immediately become eligible to participate in our Bonus Book Plan, with savings up to 70% off the publishers' prices.

4. Convenient Service. At 3-4 week intervals (16 times per year) you will receive the Book Club News, describing the Main Selection and Alternate Selections, together with a dated reply card. If you want the Main Selection, do nothing and it will be sent to you automatically. If you prefer another selection, or no book at all, simply indicate your choice on the card, and return it by the date specified. You will have at least 10 days to decide. If, because of late mail delivery of the News, you should receive a book you do not want, we guarantee return postage.

If the reply card has been removed, please write to the **Small Computer Book Club Dept. Y-AB5, Riverside, N.J. 08075** to obtain membership information and an application.



Established 1845 SCIENTIFIC AMERICAN April 1983 Volume 248 Number 4

A Nuclear-Weapon-Free Zone in Europe

A proposal to ban nuclear weapons of all kinds from a strip of land on both sides of the East-West boundary in central Europe is viewed as a politically feasible approach to reducing the risk of nuclear war

by Barry M. Blechman and Mark R. Moore

The nuclear-arms-control talks now under way in Geneva between the U.S. and the U.S.S.R. are directed toward two goals: reducing the number of strategic, or long-range, nuclear weapons deployed by the two superpowers and reducing the number of intermediate-range nuclear weapons capable of reaching targets in Europe. Excluded from the negotiations is another major category of nuclear arms, namely tactical, or short-range, nuclear weapons. Of the roughly 6,000 nuclear weapons currently deployed in Europe by the military forces of the North Atlantic Treaty Organization (NATO) more than half are classified as tactical. The opposing forces of the Warsaw Pact are thought to have a comparable arsenal of nuclear weapons based on their side of the East-West boundary in central Europe. The short-range nuclear weapons deployed by both sides in Europe are intended to provide direct support to ground forces engaged in combat; hence they are also known as battlefield nuclear weapons. Included under this heading are nuclear warheads for short-range ballistic missiles (defined here as those with a maximum range of 150 kilometers), artillery shells (with a maximum range of 30 kilometers), antiaircraft missiles and explosive mines.

The large numbers of battlefield nuclear weapons based in Europe present a special problem in arms control. Be-

Editor's Note

The Independent Commission on Disarmament and Security Issues, which included the call for a battlefield-nuclear-weapon-free zone in its final report, published last June, was made up of 18 members, all of whom served in a private capacity. The chairman of the commission was Olof Palme, then a member of the Swedish parliament, chairman of the Swedish Social Democratic Party and former prime minister of Sweden. (Palme has since been reelected Sweden's prime minister.) The other commissioners were Giorgi Arbatov, full member of the Central Committee of the Communist Party of the U.S.S.R., deputy of the Supreme Soviet and director of the Institute of the U.S.A. and Canada, a division of the Academy of Sciences; Egon Bahr, member of the West German parliament, chairman of that body's subcommittee on disarmament and arms control and former minister for economic cooperation; Gro Harlem Brundtland, member of the Norwegian parliament, chairman of the Norwegian Labor Party and former prime minister of Norway; Jozef Cyrankiewicz, former prime minister of Poland and former president of the Council of State; Jean-Marie Daillet, member of the French parliament, vice-chairman of that body's defense committee and chairman of the Giscardist UDF party's defense committee; Robert A. D. Ford, special adviser on East-West relations to the Canadian government, former Canadian ambassador to Colombia, Yugoslavia, Egypt and the U.S.S.R.; Alfonso Garcia-Robles, ambassador and chairman of the Mexican delegation to the United Nations Committee on Disarmament and former foreign minister of Mexico (co-winner of the 1982 Nobel peace prize for his role in the creation of a Latin-American nuclear-weapon-free zone); Haruki Mori, former Japanese ambassador to Britain and to the Organization for Economic Cooperation and Development and former vice-minister in the Japanese Ministry of Foreign Affairs; C. B. Muthamma, Indian ambassador to the Netherlands and former ambassador to Ghana and Hungary; Olusegun Obasanjo, general in the Nigerian armed forces, member of the Council of State, Distinguished Fellow of the University of Ibadan and former head of state of Nigeria; David Owen, member of the British parliament and former secretary of state for foreign and Commonwealth affairs; Shridath Ramphal, secretary general of the Commonwealth and former foreign minister of Guyana; Salim Salim, minister of foreign affairs of Tanzania; Soedjatmoko, rector of the United Nations University in Tokyo and former Indonesian ambassador to the U.S.; Joop den Uyl, member of the Dutch parliament, deputy prime minister of the Netherlands, leader of the Dutch Labor Party and former prime minister; Cyrus Vance, former secretary of state of the U.S.

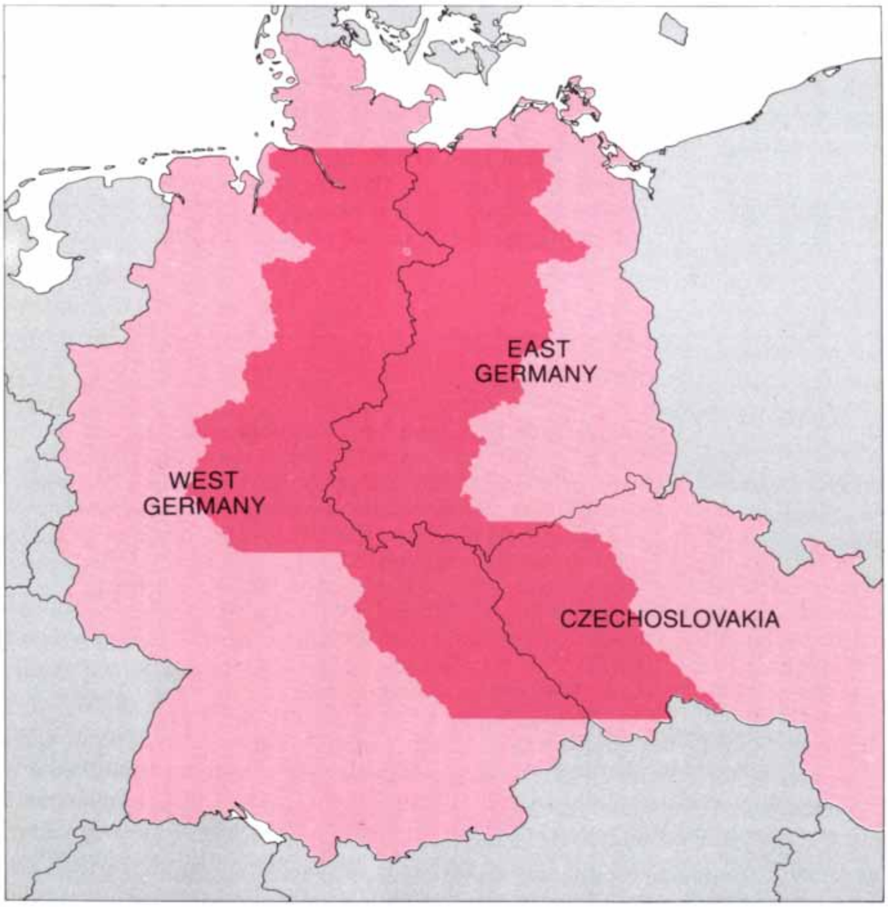
cause many of them are deployed in forward positions in central Europe, close to the front lines in any likely war, they stand a good chance of being overrun by invading forces. Given a choice between using the weapons or losing them, the defending side would presumably be under great pressure to fire them first, thereby escalating the conflict past the "nuclear threshold."

For that matter, in the kind of intense political crisis that would be likely to precede the outbreak of hostilities there could be considerable pressure on political leaders to delegate in advance to military commanders in Europe the authority to release battlefield nuclear weapons for use under certain circumstances. Whether any American president or European leader would in fact make such a decision is questionable; nevertheless, the pressure to do so would be real and the possibility cannot be dismissed. Even leaving aside the possibility of such a deliberate decision, the problem of maintaining effective control of battlefield nuclear weapons would be difficult in wartime. Given the confusion and misinformation that inevitably accompanies large-scale military operations, the specific location of nuclear-weapon sites could be a crucial factor in determining whether or not a decision is made to initiate nuclear war.

A novel plan for dealing with this problem was put forward last year by the Independent Commission on Disarmament and Security Issues, an international group of present and former government officials formed in 1980 under the chairmanship of Olof Palme, who was then the former prime minister of Sweden. (He is now again prime minister.) The commission, the members of which were from Eastern, Western and Third World nations, included Cyrus Vance, former secretary of state of the U.S., and Giorgi Arbatov, a member of the Central Committee of the Communist Party of the U.S.S.R. In a report published in June of last year the commission called for the establishment of a "battlefield-nuclear-weapon-free zone" in Europe, together with other "measures to strengthen the nuclear threshold and reduce pressures for the early use of nuclear weapons."

According to the commission's report, *Common Security: A Blueprint for Survival,* the proposed nuclear-weapon-free zone would start with central Europe, perhaps as a strip 150 kilometers wide on each side of the border of West Germany with East Germany and Czechoslovakia. In time the zone would be extended, reaching ultimately from the northern to the southern flanks of the two alliances. Nuclear weapons of any kind would be banned in the zone.

The concept of a nuclear-weapon-free zone in Europe is not new. Nevertheless, the commission's proposal differs from all the previous proposals in several important respects, not the least of which is that it is based on the characteristics of a particular category of nuclear weapons rather than on some other criterion, such as the territorial extent of the nations involved. In our view the commission's plan is a practical and politically feasible way of immediately reducing the risk of nuclear war in Europe. In this article we shall summarize the arguments for a battlefield-nuclear-weapon-free zone in central Europe. In particular we shall explain why we believe this particular plan could succeed where all the others have failed.



PROPOSED DENUCLEARIZED ZONE in central Europe would be rid of all nuclear weapons, thereby reducing the risk that in the event of a conventional military attack the defending side could be pressured into firing its battlefield nuclear weapons first in order to keep them from being overrun by the invading forces. According to the report of the Palme commission, the exact geographic definition of the zone "should be determined through negotiations" between the interested parties. For illustrative purposes, however, the commission suggested a zone consisting initially of a strip of land 150 kilometers wide on each side of the border of West Germany with East Germany and Czechoslovakia (*dark color*). Unlike earlier proposals of this type, nuclear weapons could still be stationed in parts of all three countries (*light color*).

Most nations have already agreed that certain places should remain free of nuclear weapons: Antarctica, outer space and celestial bodies, the seabed and Latin America are all subject to treaties that forbid the deployment of nuclear weapons within their domain. Proposals have also been made to ban nuclear weapons in the Middle East, South Asia, the South Pacific and Africa, and in Europe as well. Each of these proposals has been the subject of at least some discussion in various international forums. Nevertheless, the history of efforts to create nuclear-weapon-free zones has yielded far more proposals than results. Indeed, the proposals often appear to be diplomatic maneuvers for political advantage rather than genuine attempts to reduce the risk of nuclear war.

The first and most detailed plan for a nuclear-weapon-free zone in Europe was put forward in 1957 by Adam Rapacki, then foreign minister of Poland. The Rapacki plan, as it came to be known, proposed that the governments of Poland, Czechoslovakia, East Germany and West Germany agree not to manufacture, maintain or allow the installation of nuclear weapons within their borders. The denuclearized status of the zone would be guaranteed by the existing nuclear-weapons states (then the U.S., the U.S.S.R. and Britain), and a system of ground and air surveillance would be set up to ensure compliance. As the Rapacki plan was originally proposed, it did not envision a formal treaty establishing the zone; it relied instead on unilateral declarations by the interested governments.

The political objectives of the Rapacki plan seemed obvious to the Western nations. The plan was proposed at a time when the governments of the NATO countries were actively discussing the possibility of offsetting the superior conventional (non-nuclear) military forces of the Warsaw Pact by means of a "nuclear shield." Indeed, the decision to take this approach was made soon afterward, at a North Atlantic Council meeting in December, 1957. As a result the NATO countries viewed the Rapacki plan as an attempt to weaken their military position in central Europe, and they rejected it.

The Polish plan was twice revised and resubmitted to the Western nations, in November, 1958, and March, 1962. The changes did not meet NATO's basic objections, and so they made little headway. Subsequent Eastern proposals for nuclear-weapon-free zones in Europe have concentrated on subregions of the continent. The Balkans and the Adriatic area have long been the subject of proposals, and so has the entire Mediterranean basin. The Nordic region, which Premier Khrushchev of the U.S.S.R. first suggested (in 1959) as a suitable zone to be denuclearized, has in recent years again received much attention from Russian officials.

Not long after Eastern proposals for the creation of nuclear-weapon-free zones were first being made, the U.S. developed a list of four criteria that such an initiative would have to meet in order to gain American support: (1) The proposal should be initiated voluntarily by states in the region; (2) the proposal should include all the states of the region, or at least all those with significant military power; (3) the existing military balance and security arrangements should not be disturbed; (4) the agreement should include provisions for the verification of compliance.

In the case of Latin America the U.S. evidently found that the proposal to create a nuclear-weapon-free zone fitted, or could be modified to fit, these criteria. On the other hand, the U.S. and other Western governments have not found past proposals for the creation of nuclear-weapon-free zones in Europe to be suitable subjects for negotiation. Some of the objections have been technical: given the long-standing opposition of the U.S.S.R. to on-site inspection, the task of ensuring the withdrawal and continued absence of nuclear weapons from a region already densely populated with them has seemed difficult if not impossible.

	WEAPON	FIRST DEPLOYED	RANGE (KILOMETERS)	MAXIMUM YIELD (KILOTONS)	NUMBER OF LAUNCHERS
NATO	LANCE	1973	115	50	90
	HONEST JOHN	1953	38	20	90
	M109 155-MILLIMETER GUN	1964	18	2	1,455
	M110 203-MILLIMETER GUN	1962	16	1	390
	PLUTON (FRANCE)	1974	120	10	42
WARSAW PACT	FROG/SS-21	1950's/1977	75/155	200	375
	SCUD-A	1957	85	1	250
	203-MILLIMETER GUN	1979	31	?	150
	240-MILLIMETER GUN	1979	30+	?	150

BATTLEFIELD NUCLEAR WEAPONS currently deployed on both sides of the East-West boundary in central Europe include the short-range ballistic missiles and artillery pieces listed in this table. Included for the forces of the North Atlantic Treaty Organization (NATO) are all such systems assigned to the European theater of operations. (France, although it is a member of NATO, is not an integrated part of the military alliance.) Included for the forces of the Warsaw Pact are all such systems stationed outside the U.S.S.R. and in the western military districts of the U.S.S.R. All the missiles shown are capable of delivering only one nuclear warhead each. Not shown are nuclear explosive mines and nuclear-armed surface-to-air missiles, both of which are sometimes classified as tactical nuclear weapons. Also excluded are the thousands of longer-range nuclear weapons deployed on land or at sea within range of European targets.

The fundamental Western objections to nuclear-weapon-free zones in Europe, however, have arisen from two other factors: the basic asymmetry in the geopolitical characteristics of the U.S. and the U.S.S.R., and the key role nuclear weapons have come to play in Western military strategy and in political relations among the NATO partners. Given the weight of these considerations, Western officials have tended to see Eastern proposals for the creation of nuclear-weapon-free zones in various parts of Europe as being largely propagandistic, meant to encourage dissension within the Western European nations and to disrupt political bonds among the members of the alliance. NATO's reservations must be fully understood if realistic measures to reduce the risk of nuclear war in Europe are to be designed.

Throughout the history of the Western alliance officials and ordinary citizens in the Western nations have debated the degree to which it would be effective and prudent to rely on nuclear weapons to maintain the peace in Europe. Advocates of conventional defenses argue that the surest route to stability in Europe is to maintain conventional military forces strong enough to persuade the leaders of the U.S.S.R. that in the event of conflict they would be defeated on the battlefield and thus face a choice between escalating to nuclear war or accepting defeat. If the Russian leaders believed such a choice would be the likely result of conflict in Europe, it is argued, they would decide against war, choosing instead to favor stable relations between the Eastern and Western nations. Support for this position has led to recurrent efforts to strengthen NATO's conventional forces.

For many other people in the Western nations, however, the threat of nuclear war, and the forces necessary to make the threat credible, have always been the essence of NATO's strategy. From this point of view what persuades the leaders of the U.S.S.R. to seek stability and avoid war in Europe is not a calculation that they would be likely to lose such a conflict but rather a general perception that if events were to get out of hand, the results of any military engagement could be disastrous. Indeed, people holding this view argue that NATO's military forces should be designed to avoid a situation in which the results of conflict become calculable. If the outcome can be predicted, they maintain, it may become acceptable under certain conditions. It would be far better, they argue, to enhance deterrence by inculcating the overriding perception that the inexorable result of conflict would be catastrophe.

Those holding this view seek to make clear to the leaders of the U.S.S.R. that any overt aggressive act in Europe could easily deteriorate into conflict and eventually escalate into an all-out nuclear war between the U.S. and the U.S.S.R. In short, they seek to maintain a position in which the ultimate danger of confrontation in Europe—nuclear holocaust—is made as stark as possible. No improvement in conventional forces can serve this purpose. Indeed, from this point of view the stronger conventional defenses seem, the greater the risk is that the leaders of the U.S.S.R. might err in their calculations and commit aggression, in the belief that such defenses had reduced the risk of nuclear war.

For this reason, among others, NATO has never fulfilled its stated objective of strengthening its conventional forces. It is also the reason NATO declares explicitly that, if necessary, it would initiate and escalate a nuclear war in Europe. Furthermore, it is the reason NATO maintains thousands of tactical nuclear weapons in Western Europe, is now deploying a new generation of intermediate-range nuclear weapons and regularly expresses concern about the balance between the strategic nuclear forces of the U.S. and of the U.S.S.R.

There is yet another reason, however, for this emphasis on nuclear defense. For almost 40 years American strategists and political leaders have believed the most dangerous threat to American interests is represented by the U.S.S.R. The U.S.S.R. is perceived not only as threatening U.S. security in military terms but also as challenging in subtler ways the values of American society. As a result, when American strategy is viewed in geopolitical terms, its essence throughout this period has been to contain both the military power and the political influence of the U.S.S.R. within the Eurasian land mass. Most important, U.S. strategy has sought to limit the influence of the U.S.S.R. in the parts of Europe beyond the belt of nations occupied by the Russian armies in 1945. The chosen vehicle of the strategy, of course, is NATO.

In the struggle for influence in Western Europe between the two superpowers the U.S. is at a distinct geographic disadvantage. The military power of the U.S.S.R. is an immutable presence felt strongly in the capitals of Western Europe. In contrast, regardless of the vast economic and political stake of the U.S. in the continued independence and pro-American orientation of Western Europe, the U.S. security commitment and military presence in Europe is now, and always will be, perceived as somewhat transient and unreliable. Western Europeans recognize that the U.S. always has the option of withdrawing beyond the oceans that have sheltered it in the past. Such Western European fears can never be fully allayed. Indeed, they often are aggravated when the U.S. shows evidence of having one of its recurrent bouts of neoisolationism.

To offset the asymmetry in both the perception and the reality of the American and Russian military presence in Western Europe, U.S. decision makers have made continual efforts to strengthen the links between U.S. forces stationed in Western Europe and overall U.S. military power, and to reassure allies (and caution adversaries) about the durability of U.S. commitments. These efforts take various forms, but the crucial components have been steps to "couple" U.S. strategic nuclear forces with ground forces deployed in Europe.

Western strategy in Europe is based on the concept of escalated responses to aggression, beginning with conventional counterattacks if the Warsaw Pact attack is limited to conventional forces but including the possibility of the initiation and continued escalation of nuclear war. This declared position is given weight by the deployment of short-range nuclear weapons in Europe not far from the putative front lines; these weapons, which are designed to be used against military forces on the battlefield, would in many scenarios be the first to be used if a decision were made to cross the nuclear threshold. Because they are of comparatively small yield and would be employed against military targets, they are believed to make the threat of initiating nuclear war more credible.

To add further credibility to NATO's nuclear threat, the alliance plans to deploy U.S. intermediate-range missiles in Western Europe that are capable of striking targets in Russian territory. If the use of battlefield nuclear weapons did not lead to a termination of the conflict, then the intermediate-range weapons could be launched, still against military targets. This step would presumably be less difficult than the next one: the engagement of the central strategic forces of the U.S. Even so, nuclear strikes on Russian territory by U.S.-controlled missiles would almost certainly lead to a Russian response against the U.S., thereby precipitating the involvement of the two sides' central strategic forces. The idea is that since the leaders of the U.S.S.R. would recognize this sequence of escalation, they would be deterred from taking the first step on the escalation ladder. In theory deterrence is strengthened to the extent that the eventual involvement of U.S. strategic nuclear forces is made credible.

In one sense this sharing of risk is the ultimate psychological and practical bond of the NATO alliance, at least with respect to its security policies. By promising to resort, if necessary, to strategic nuclear forces in defense of Western Europe, thereby accepting the likelihood of retaliatory Russian attacks on U.S. territory, the U.S. has injected itself symbolically as a full partner in the fate of Europe, thereby offsetting the geographic isolation of the U.S.

In the light of this reasoning, the flaws in Eastern proposals to denuclearize parts of Europe, as seen in NATO capitals, should be apparent. As traditionally proposed, a nuclear-weapon-free zone would deprive Western Europe of a vital link between its conventional defenses and the U.S. strategic nuclear deterrent, thus emphasizing existing asymmetries both in the two sides' geographic positions and in the balance of conventional military forces. This reasoning applies primarily to proposals (such as the Rapacki plan) for the creation of nuclear-weapon-free zones in central Europe, the subregion at the heart of all considerations of the defense of Western Europe. Similar proposals for other subregions share some of the

U.S. LANCE MISSILE is seen being readied for a test launch from its self-propelled erector-launcher vehicle. The liquid-fuel rocket can carry either a nuclear warhead or a conventional one. Its maximum range is about 115 kilometers. The Lance has replaced the Honest John as the standard U.S. battlefield nuclear missile. Other NATO nations have both the Lance and the Honest John in their arsenals. All nuclear warheads for the missiles are in U.S. custody.

U.S. M110 203-MILLIMETER GUN is shown being test-fired with a non-nuclear shell. The eight-inch howitzer is also capable of firing a nuclear shell. Its maximum range is approximately 16 kilometers. The gun can be mounted either on a self-propelled vehicle, as is seen here, or on a towed chassis. The structure at the rear absorbs the recoil. The U.S. has supplied M110's to most of its NATO allies.

same political and military deficiencies, however, and in any case they are regarded as unhelpful precedents that might make a central European nuclear-weapon-free zone harder to reject.

From NATO's standpoint, therefore, the weaknesses of nuclear-weapon-free zones, as they have been traditionally proposed, are as follows. First, the initiatives typically are not coupled with proposals for restraints on conventional forces. This means that the ability of NATO to offset the Warsaw Pact's greater conventional forces with the threat of nuclear escalation would be undermined. Of course, NATO could mobilize its greater economic resources to compete more effectively in terms of conventional military forces, but that would be expensive and seems unlikely in the face of the traditional reluctance of the Western democracies to sustain heavy military burdens in peacetime.

Moreover, even if more effective conventional defenses were to be mounted, there would remain the question of the extent to which the lessening of NATO's nuclear threat would reduce the ability of the Western European nations to resist Russian political pressures and even military aggression. Although it would be impossible to prove, the fact that international relations in Europe have evolved into a fairly stable configuration over the past 35 years suggests that NATO strategy may well have fulfilled its designers' military and political expectations. At a minimum it indicates the existing doctrine should not be modified without substantial reason to expect that alternatives would be equally successful and better suited to current circumstances.

Second, because the removal of nuclear forces from central Europe would suggest the decoupling of the American strategic nuclear deterrent from the defenses of Western Europe, it might be taken to presage the political decoupling of Western Europe from the U.S. Such a development, it is feared, would lead inevitably to greater Russian influence in, and eventual dominance over, European affairs. The fact that Russian territory was not included in proposals to denuclearize central Europe highlighted the problem. It meant in effect that if NATO accepted Rapacki's proposal or subsequent ones, it would be acquiescing in a diminution of the American presence in Western Europe without any commensurate diminution of the overpowering Russian presence. The political implications of such a poor bargain are striking. It would be widely interpreted as suggesting that decisive Russian influence in the affairs of Western Europe had been conceded.

One more concern enters NATO evaluations. In addition to its purpose of containing Russian power, the formal partnership of West Germany with other Western European and North American states in a military alliance was intended to help heal the wounds of World War II. Along with the various multinational economic institutions created in the 1950's, NATO fostered the reconciliation of Germany and its former enemies, making possible the development of stable relations in Europe and the avoidance of the national tensions that had led to disastrous wars three times in the span of a century.

The Rapacki plan would have placed East and West Germany (along with Poland and Czechoslovakia) in a special category. West Germany would no longer have been a full member of the alliance. Moreover, such a decision would not have been made by the Germans themselves but rather would have been imposed on Germany by outside forces. This would have emphasized Germany's special position and, it was feared, would have strengthened prospects for the emergence of nationalist forces within Germany that could jeopardize its political ties with other Western nations and eventually those between the Eastern and Western nations. In short, by singling out Germany, the Rapacki plan and other proposals for nuclear-weapon-free zones in central Europe raised fears that old tensions might be rekindled. Far better, in the eyes of many Western leaders, would be the continued integration of Germany with the rest of Western Europe and the gradual improvement of relations between the two alliances as a whole.

Thus the Rapacki plan and subsequent proposals for nuclear-weapon-free zones in central Europe were marred primarily by their political implications. Because they suggested a clear differentiation of risk between the U.S. and Europe, they were interpreted as being means for the erosion of American influence in Europe and the aggrandizement of Russian power. Because they singled out Germany for special treatment, they were perceived as being likely to undermine the efforts that had been made to integrate West Germany into the Western community of nations. The current proposal of the International Commission on Disarmament and Security Issues for a battlefield-nuclear-weapon-free zone has been carefully designed to avoid these disadvantages.

The commission took careful note of the complex relations among the different types of military forces in Europe and urged a comprehensive approach to arms control as the only way to ensure the achievement of lower levels of forces and a reduced risk of war. The commission stated in its report that "a negotiated agreement for approximate parity in conventional forces between the two alliances... would facilitate reductions in nuclear weapons and a reordering of the priority now accorded to nuclear arms in military contingency planning."

Accordingly the commission began its recommendations for Europe by urging the convening of a conference of the foreign ministers of the NATO and Warsaw Pact nations that have been negotiating in Vienna for the past 10 years for an agreement on conventional forces. It called on the foreign ministers to complete a first-phase agreement that would guarantee parity between the two sides and make possible continued negotiations for substantial reductions in conventional forces. In the view of the members of the commission, the differences remaining between the two sides could be resolved fairly quickly if there were the political will to resolve them. The commission concluded that agreement to hold a meeting at the foreign-minister level could provide both the impetus and the opportunity to move things forward.

The commission also paid considerable attention to chemical weapons, urging the removal of such weapons from all Europe and proposing associated measures to assure both sides that such a prohibition was being observed. The latter measures would include on-site inspections on a "challenge" basis, meaning that each side would be allowed to inspect a certain number of facilities of its choosing on the other's territory each year, if it wanted to exercise the option. This is an important and creative proposal. That the Russian representative on the commission agreed to the inspection provision suggests a greater Russian responsiveness to Western demands for serious attention to the problems of verifying arms-control agreements than has been evident in the past. Russian responsiveness to concerns about verification was also evident in the discussion of the proposed nuclear-weapon-free zone.

The commission considered several alternative approaches to the nuclear component of European security problems, including the creation of nuclear-weapon-free zones defined by national boundaries. In the end, however, the members of the commission decided to pursue a functional approach, concentrating on specific categories of weapons that raise particular problems.

The commission called special attention to "the dangers posed by those nuclear weapons whose delivery systems are deployed in considerable numbers to forward positions in Europe." It urged the mutual removal of battlefield nuclear weapons from regions close to the East-West border through the negotiation of a battlefield-nuclear-weapon-free zone. To minimize verification difficulties the commission proposed that no nuclear munitions be allowed in the zone. It stated further that storage sites for nuclear munitions, maneuvers simulating nuclear operations and preparations for the emplacement of nuclear mines should also be prohibited. Finally, it suggested special rules should be negotiated covering the deployment in the zone of "dual purpose" artillery and short-range missiles that could be used to deliver nuclear munitions.

The commission believed it was best not to recommend specific boundaries for the zone, leaving that task to the parties concerned. For illustrative purposes, however, the report cited the possibility of a 300-kilometer zone centered on the East-West boundary [*see illustration on page 38*]. Such a zone would initially include parts of West Germany, East Germany and Czechoslovakia. Unlike the Rapacki plan and subsequent Eastern proposals, however, parts of all three nations would still be available for the deployment of nuclear weapons; this was a crucial distinction. As the zone was extended to the north and the south, depending on how its boundaries were set, parts of Austria, Hungary, Yugo-

RUSSIAN SHORT-RANGE MISSILES are transported across a muddy field on tracked erector-launcher vehicles. The missiles, designated SCUD-A by Western intelligence sources, can fire either a nuclear warhead or a conventional one. Their maximum range is

slavia, Italy, Denmark, Sweden, Norway, Finland and the U.S.S.R. could be added to it.

The commission also left it to the negotiators to define the specific provisions needed to ensure the parties' ability to verify compliance with the agreement. The members did agree, however, that such provisions "would have to include a limited number of on-site inspections in the zone on a challenge basis." Again the fact that the Russian representative agreed to include this phrase is significant. Although Arbatov, like all the other members of the commission, served in a private capacity, it is evident that his assent to the text—particularly the recommendations—was not given without higher authorization. The acceptance of on-site inspections in this context is consonant with the trend toward greater Russian receptivity to more intrusive means of verification.

Finally, the commission recommended some additional measures to further strengthen the nuclear threshold and reduce pressures for the early use of nuclear weapons. Most pointed was the recommendation of a majority of the commission that nuclear weapons that might appear more usable, particularly small nuclear weapons such as enhanced radiation weapons ("neutron bombs"), not be deployed in Europe. Although a specific recommendation was not made, a section of the report also discussed the special problems associated with nuclear-armed antiaircraft missiles and explosive mines, suggesting that the removal of these weapons from Europe could further strengthen the barriers against the initiation of nuclear war.

about 85 kilometers. Nuclear warheads for Warsaw Pact weapons are in Russian custody.

In short, the battlefield-nuclear-weapon-free zone proposed by the commission appears to escape the flaws associated with previously proposed nuclear-weapon-free zones. What would the implementation of the battlefield-nuclear-weapon-free zone accomplish?

For one thing, it would strengthen the barriers against the inadvertent or accidental initiation of nuclear war. It seems clear that nuclear war is not likely to result from a cold-blooded calculation of advantage during normal times; the forces of both sides are too large for either side to see any meaningful advantage to be gained. It is likelier that nuclear war would emerge from desperate decisions in an intense crisis, probably one in which the conventional forces of the two sides had already been engaged. Under such circumstances one side or the other might initiate a nuclear attack because it saw no other option for averting a catastrophic defeat, or because it had concluded that its adversary was about to escalate to the nuclear level.

The removal of all nuclear weapons from a zone of substantial width on both sides of the East-West boundary in central Europe would reduce the chances of the contingency most likely to precipitate such desperate decisions. It would remove these weapons from the chaos of warfare, at least in the early days. It would give leaders on both sides a most precious commodity: time. It would allow more careful and considered analysis of the tactical military situation and the possibility of persevering through the continued use of conventional forces only. It would reduce the concern on both sides that the other side would feel pressure to release its weapons hastily, and it would correspondingly lessen the urge to launch a preemptive attack. Finally, it would allow the greater exploration of diplomatic alternatives; during the increased time available before a decision to release nuclear weapons would have to be made, both sides might come to see the extraordinary dangers of the course on which they had embarked and take steps to resolve the conflict before it got out of hand.

There would be other advantages to implementing the commission's proposal for a battlefield-nuclear-weapon-free zone. The changes necessitated by such an agreement might lead both sides independently to reconsider the current inventories of nuclear weapons on the continent. Many observers have argued that substantial reductions are possible on military grounds. What has forestalled them are concerns about the political implications of withdrawing weapons unilaterally. The relocations made necessary by the battlefield-nuclear-weapon-free zone, and the fact that the other side was making similar changes, might provide the opportunity for substantially reducing nuclear forces without suffering adverse political consequences.

The successful negotiation of a battlefield-nuclear-weapon-free zone, along with the proposed measures to restrain conventional and chemical weapons, would open more ambitious possibilities for controlling and reducing military arsenals in Europe and elsewhere; it would have a positive effect on political relations between the Eastern and Western nations, implying the start of progress toward less tense and more cooperative relations between the U.S. and the U.S.S.R., and it would be good for relations within the Western alliance as well, eliminating what has become a damaging source of friction between the U.S. and Western Europe: the debate over approaches to nuclear-arms control on the Continent.

Late last year the Swedish government formally solicited the views of all European and North American governments on the proposed battlefield-nuclear-weapon-free zone. In response the U.S.S.R. stated in January that it endorsed the idea and would put such a proposal on the table at the Vienna talks. It also said, however, that the present proposal did not go far enough, adding that the zone should be between 500 and 600 kilometers wide. Such an expansion would take in virtually all of both East Germany and West Germany.

So far the Western nations have remained cool to the idea of a denuclearized zone. If the Russians do indeed put forward such a proposal in Vienna, however, there would be a golden opportunity to link the implementation of a denuclearized zone explicitly to the conclusion of a first-phase agreement on the limitation of conventional forces, and to negotiate a mutually acceptable width for the zone that did not give rise to political problems similar to those presented by the Rapacki plan.

The implementation of a battlefield-nuclear-weapon-free zone would not eliminate the risk of nuclear war in Europe. Some risk will remain as long as nuclear weapons remain in the inventories of the major powers. The degree of risk will wax or wane in consonance with overall political relations and specific changes in military deployments on the two sides, but it is unlikely ever to disappear. It is a fact with which everyone will have to continue to live. The proposed battlefield-nuclear-weapon-free zone, however, is a pragmatic and politically feasible measure to control the danger of nuclear war in Europe. It merits serious consideration by all citizens and governments.

Silicon Micromechanical Devices

Tiny valves, nozzles, pressure sensors and other mechanical systems can be chemically etched in a wafer of single-crystal silicon. Such devices can be mass-produced much as microelectronic circuits are

by James B. Angell, Stephen C. Terry and Phillip W. Barth

In the past 30 years the element silicon has become familiar as the material from which electronic components and systems are made. In the form of microelectronic chips it can be found in everything from dishwashers to the Space Shuttle control systems. The exploitation of silicon's electrical properties has been accompanied by a less publicized exploration of its other properties and their potential uses. This research has led to the development of a technology called micromachining that allows silicon to be made into mechanical devices almost as small as microelectronic ones.

Micromachining starts with the same batch-fabrication techniques that have made silicon integrated-circuit chips inexpensive. The techniques make it possible to fabricate many chips at once, so that the cost of production is spread over many chips and the cost per chip is low. To this repertory of techniques micromachining adds chemical etching techniques for forming three-dimensional shapes such as pits, holes, pyramids, trenches, hemispheres, cantilevers, diaphragms, needles and walls. A wide variety of mechanical devices can be constructed from combinations of these structural elements.

Among the silicon micromechanisms that have been built are valves, springs, mirrors, nozzles, connectors, printer heads, circuit boards, heat sinks and sensors for properties such as force, pressure, acceleration and chemical concentration. Even a device as complex as a gas chromatograph, an instrument for identifying and measuring gases in an unknown mixture, can be built on a disk of silicon a few centimeters in diameter.

Much of the interest in micromachining derives from the need for cheaper and more versatile sensors. Until recently the electronic components of automatic-control systems and measuring instruments were more expensive than the sensors. With the advent of the microprocessor the cost of electronics began to fall dramatically. Batch-fabricated silicon sensors could replace expensive hand-assembled sensors much as batch-fabricated silicon circuits have replaced vacuum-tube assemblies.

The latest development in this area is the integrated sensor, a silicon chip that includes both a sensor and associated signal-conditioning electronics. Integrated sensors are less expensive than sensors and the necessary electronic components fabricated separately. In addition, because the signal from an integrated sensor is less vulnerable to noise and leakage, integrated sensors perform better than discrete sensors and have a wider variety of applications.

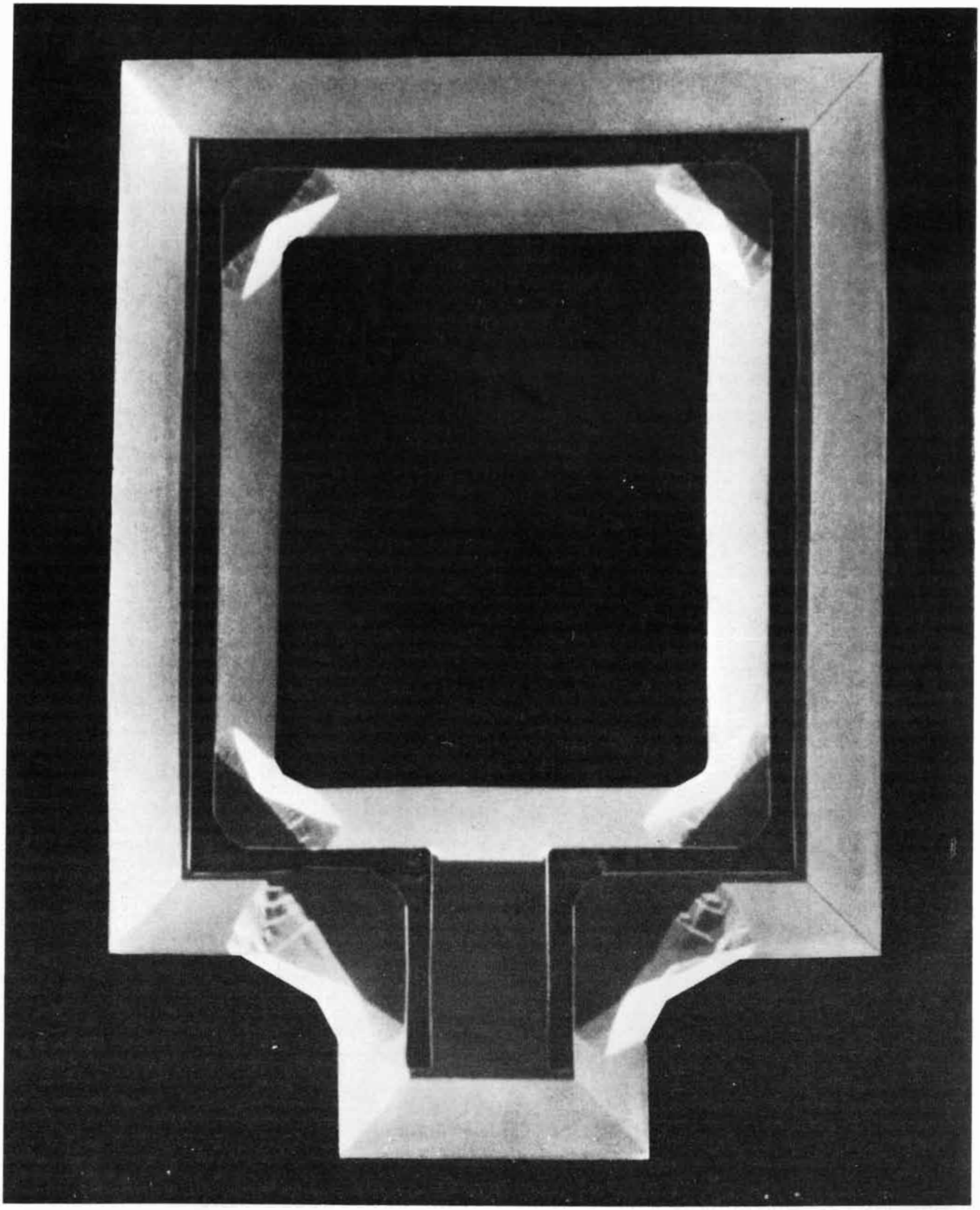
Silicon is a semiconductor, one of the elements that lie between the metals and the nonmetals in the periodic table of the elements. It is distinguished from other semiconductors, however, in that it can be readily oxidized. Silicon forms a surface layer of silicon dioxide (SiO_2) when it is exposed to steam. The layer is chemically inert and electrically insulating; essentially it is a glass. Oxide layers are used to protect areas of the silicon during the batch fabrication of microelectronic devices, which is why most such devices are made out of silicon rather than out of other semiconductors such as germanium. Micromachining also makes extensive use of silicon dioxide surface layers.

The silicon employed in the electronics industry takes the form of single crystals whose production is in itself a sophisticated technology. Large "boules," or single crystals of silicon, 10 centimeters in diameter and a meter long, are grown and sliced into disk-shaped wafers from .2 to .5 millimeter thick. The wafers are polished to a mirror finish to remove flaws created by sawing. The homogeneous crystal structure of the material gives it the electrical properties needed in microelectronic circuits. As it turns out, silicon in this form also has desirable mechanical properties.

Silicon, which is directly below carbon in the periodic table, forms the same type of crystal as diamond, although the interatomic bonds are somewhat weaker. Single-crystal silicon is brittle and can be cleaved like a diamond, but it is harder than most metals. In addition single-crystal silicon is surprisingly resistant to mechanical stress. In both tension and compression it has a higher elastic limit than steel; on the other hand, when the limit is reached, silicon fractures whereas steel deforms inelastically. Finally, as a single crystal silicon remains strong under repeated cycles of tension and compression, whereas polycrystalline metals tend to weaken and break because stresses accumulate at the intercrystal boundaries.

Both micromachining and microelectronic fabrication begin with photolithography, the photographic technique used to transfer copies of a master pattern onto the surface of a silicon wafer. The first step in photolithography is to grow a thin layer of oxide on the wafer surface by heating it to between 800 and 1,200 degrees Celsius in an atmosphere of steam. (Dry oxygen also works, but steam is much faster.) Next, a thin layer of an organic polymer that is sensitive to ultraviolet radiation, called a photoresist, is deposited on the oxide surface. A photomask, generally a glass plate on which there is a metal pattern, is placed in contact with the photoresist-coated surface, and the wafer is exposed to ultraviolet radiation. The metal on the mask is opaque to ultraviolet wavelengths, whereas the glass is transparent. The radiation causes a chemical reaction in the exposed areas of the photoresist. If the photoresist is of the type called positive, the chemical reaction weakens the polymer. If the photoresist is of the type called negative, the reaction strengthens the polymer. The wafer is rinsed in a developing solution that removes either the exposed areas or the unexposed areas of photoresist, leaving a pattern of bare and photoresist-coated oxide on the wafer surface. The photoresist pattern is either the positive or the negative image of the pattern on the photomask.

The wafer is then placed in a solution

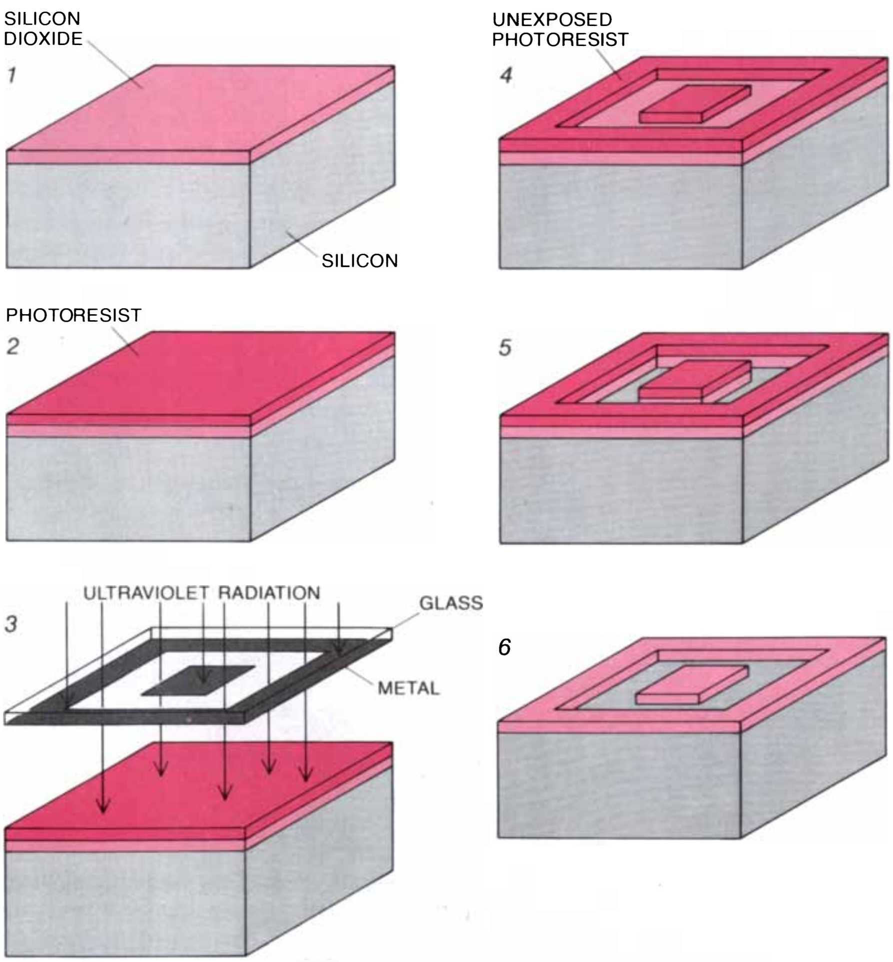


SILICON ACCELEROMETER developed by Lynn M. Roylance of Stanford University is a miniature instrument that exploits both the mechanical and the electrical properties of silicon. It is made by selective chemical etching of a "wafer" cut from a large single crystal of silicon. The accelerometer, shown here in a scanning electron micrograph, is essentially a mass of silicon with a thickness equal to that of the wafer, suspended at the end of a thin silicon beam. An electrical resistor is fabricated in the underside of the beam; its resistance changes when the beam flexes, yielding a measure of acceleration. The mass is freed from the surrounding material by etching completely through the silicon at the periphery of the mass. The beam is defined by etching away all but a thin layer of silicon in that part of the wafer. The packaged device covers an area of two millimeters by three millimeters and is .6 millimeter thick, small enough to be sutured to the heart to measure its acceleration. Hundreds of accelerometers can be fabricated simultaneously on one silicon wafer.

of hydrofluoric acid, which attacks the oxide but not the photoresist or the underlying silicon. The photoresist protects the oxide areas it covers. Once the exposed oxide has been etched away the remaining photoresist can be stripped off with hot sulfuric acid, which attacks the photoresist but not the oxide or the silicon. The final result is a pattern of oxide on the wafer surface that duplicates the photoresist pattern and is therefore either a positive or a negative copy of the pattern on the photomask. The oxide pattern itself serves as a mask in subsequent processing steps.

At this point micromachining diverges from microelectronic fabrication. In processing an electronic device the oxide pattern serves as a mask during the "doping" of the wafer with impurities such as boron or phosphorus. When the wafer is heated, the impurity atoms deposited on its surface diffuse a short distance into the silicon underlying the openings in the oxide, creating the shallow conductive or resistive regions of which microelectronic devices are composed. In micromachining the oxide is used as a mask during chemical etching. The etchants attack the silicon underlying openings in the oxide layer, excavating deep three-dimensional pits in the wafer.

Two types of etchants are employed in the micromachining of silicon. Isotropic etchants etch the silicon crystal at the same rate in all directions and create gently rounded shapes. Anisotropic etchants, which are also known as orientation-dependent or crystallographic etchants, etch at different rates in different directions in the crystal lattice; they can form well-defined shapes with sharp edges and corners. The most useful isotropic etchants are mixtures of hydrofluoric, nitric and acetic acids (HNA etchants). The anisotropic etchants are all hot alkaline solutions such as aqueous potassium hydroxide (KOH), aqueous sodium hydroxide (NaOH) and a mixture of ethylenediamine, pyrocatechol and water known as EDP.



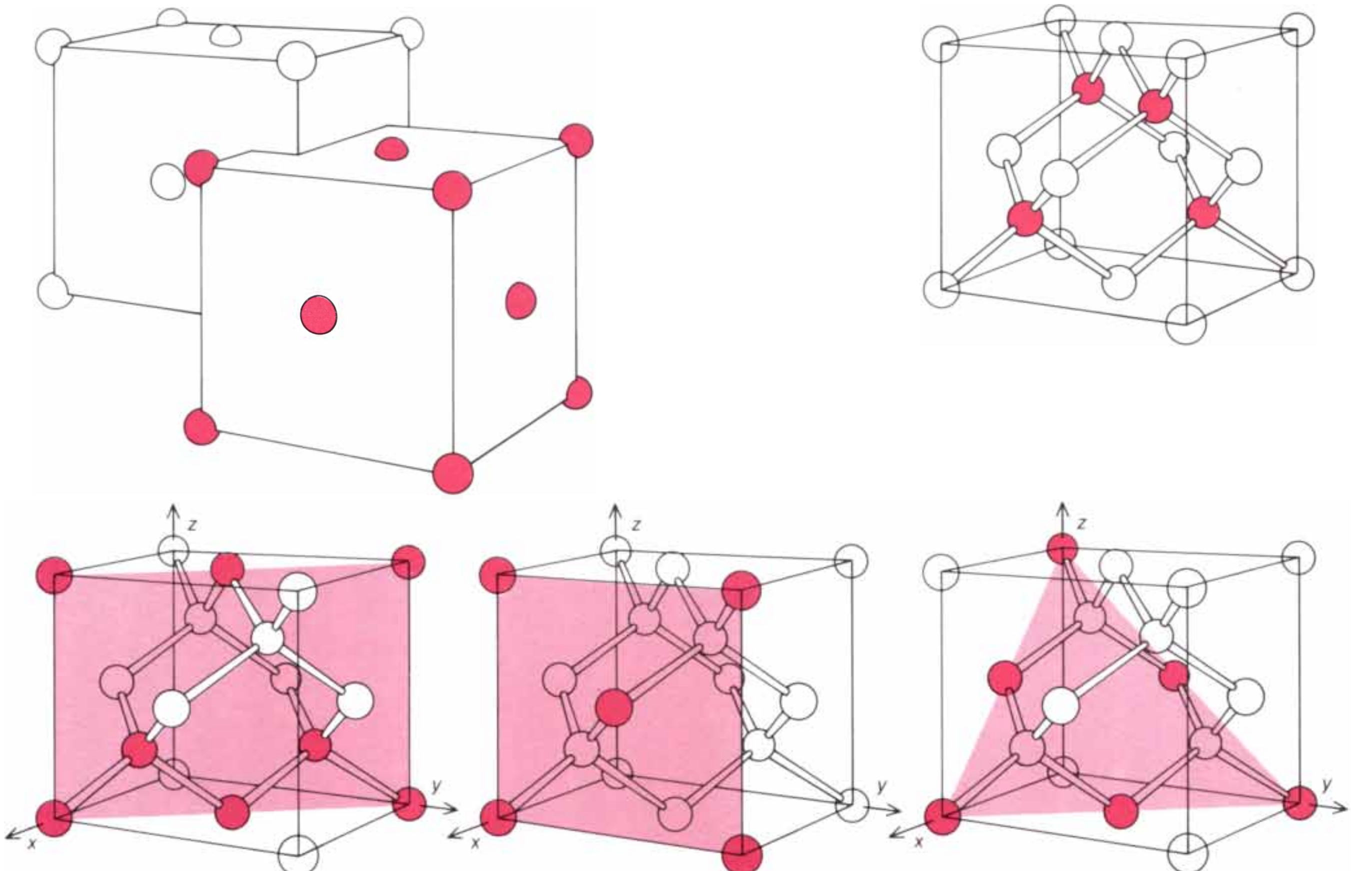
OXIDE PATTERNS on the wafer surface protect selected areas during the wet chemical etching that shapes a micromechanical device. First, a layer of silicon dioxide (SiO_2) is grown on the wafer surface; silicon reacts with steam at temperatures between 800 and 1,200 degrees Celsius to form the oxide. The oxidized wafer is coated with a polymer sensitive to ultraviolet radiation, called a photoresist. Next the wafer is placed under a master mask (a metal pattern on a glass plate) through which it is exposed to radiation. Photoresist in exposed areas weakens and is rinsed away in a developing solution. Then the wafer is etched in an acid that attacks the uncovered oxide but not the remaining photoresist or the silicon. The result is a pattern of openings in the oxide that duplicates the metal pattern on the glass plate. In subsequent processing steps the oxide pattern serves as a mask for the etching of the underlying silicon.

Micromachining makes more extensive use of the anisotropic etchants than of the isotropic ones. A typical micromachining operation, the etching of an array of holes completely through the wafer, illustrates why. An isotropic etchant moves both downward and outward from an opening in the oxide, undercutting the oxide mask and enlarging the etched pit while deepening it. The pit created is at least twice as wide as it is deep, so that the openings in the oxide must be placed at least as far apart (edge to edge) as the wafer is thick if separate holes are to be etched through the wafer. On the other hand, an anisotropic etchant, if it is properly employed, does not undercut an oxide mask and creates pits with well-defined side walls. The pits deepen without widening. Thus an anisotropic etchant can create closely spaced arrays of holes; the edges of the openings in the oxide mask can be placed as close together as the limits of photolithography allow.

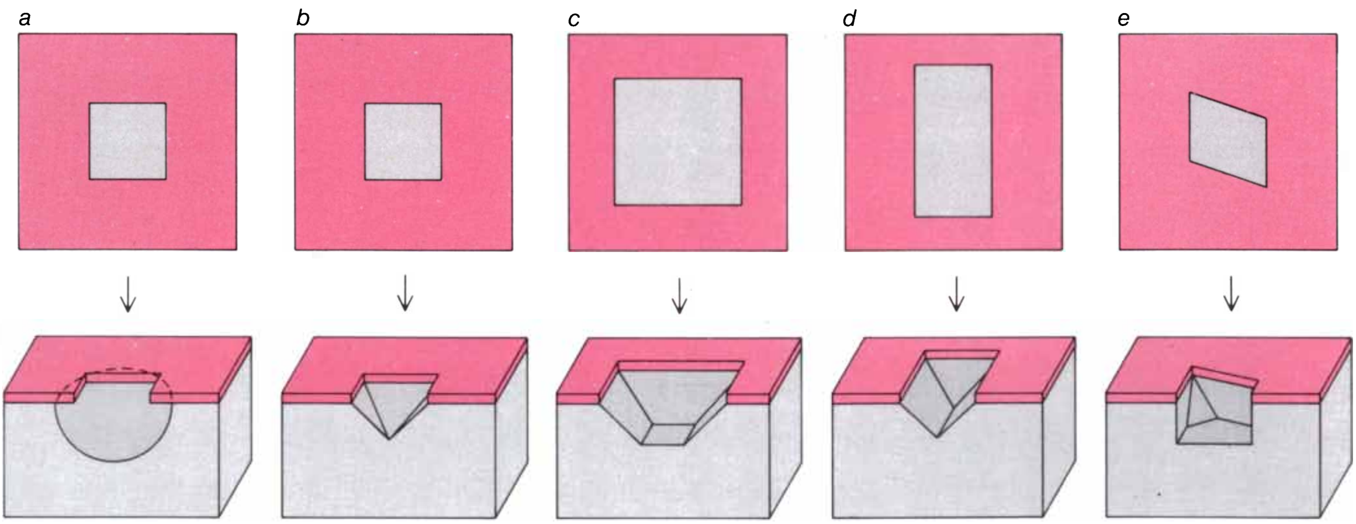
The shape of a hole formed by an anisotropic etchant depends on the orientation of the atomic planes in the silicon wafer. Just as various one-dimensional rows of corn stalks can be identified at different angles in a two-dimensional cornfield, so various two-dimensional planes of atoms lie at different angles in a three-dimensional crystal. The planes of atoms in the crystal are identified by means of a coordinate system assigned to the crystal.

Each silicon atom bonds with four neighboring atoms. The bonding structure is tetrahedral, that is, each atom lies at the center of a tetrahedron defined by the four atoms with which it is bonded. The bonding structure is an inconvenient basis for a coordinate system because the atomic bonds are not at right angles to one another. It is more convenient to think of the silicon crystal as being composed of stacked layers of repeating cubes. Each cube has an atom at each corner and at the center of each face and interlocks with four neighboring cubes. This interlocking face-centered cubic structure can serve as the basis of a coordinate system. Any atom can be designated as the point of origin of a coordinate system whose axes lie along the edges of the cube to which the atom belongs. The basic unit of measure is the length of the edge of one cube.

A vector, or direction in the crystal, is described by three coordinates known as the Miller indexes. For example, a [110] vector points diagonally across the face of a unit cube. The coordinates of a given vector also designate the entire set of atomic planes perpendicular to that direction. Thus the notation (100) de-



CRYSTALLINE SILICON shares with diamond the crystal structure called interlocking face-centered cubic. An atom lies at each corner and in the center of each face of the unit cube, and the cubes interlock in such a way that several atoms from neighboring cubes lie within each unit cube. The axes of the unit cube provide a rectilinear coordinate system that makes it possible to specify directions and planes within the crystal. A crystalline direction is designated by three coordinates called the Miller indexes, which are integer multiples of the length of one edge of a unit cube. The same set of indexes designates the planes perpendicular to the direction. Crystalline orientation is important in the fabrication of micromechanical devices because some of the etchants employed attack different directions in the crystal at different rates. Most such anisotropic etchants progress rapidly in the crystal direction perpendicular to the (110) plane and less rapidly in the direction perpendicular to the (100) plane. The direction perpendicular to the (111) plane etches very slowly, if at all.



ANISOTROPIC ETCHANTS create faceted holes composed of the crystal planes that are etched slowest, unlike isotropic etchants, which invariably produce a gently rounded hole (*a*). The shape of an anisotropically etched hole is determined by the crystalline orientation of the wafer surface, the shape and orientation of the openings in the mask on that surface and the orientation dependence of the etchant itself. A square opening oriented along the $\langle 110 \rangle$ directions of a $\langle 100 \rangle$ wafer yields a pyramidal pit with $\{111\}$ side walls (*b*). If the opening is larger, the point of intersection of the $\{111\}$ planes is deeper, and a flat-bottomed pit can be created by stopping the etch before that depth is reached (*c*). A rectangular opening on the same wafer gives a *V*-shaped groove (*d*). Holes with parallel side walls can be created by etching a wafer that has a different surface orientation. In a $\langle 110 \rangle$ wafer two sets of $\{111\}$ planes are perpendicular to the surface although not to each other. If an oxide opening on a $\langle 110 \rangle$ wafer is properly oriented, etching creates a hole with vertical side walls. The side walls that intersect at acute angles are bridged by still other planes. These are just a few of the many shapes that can be etched.

scribes all the planes perpendicular to the *x* axis. Because the crystal structure is symmetrical, the *x*, *y* and *z* directions are interchangeable, and a generalized notation can be introduced to describe all equivalent directions and sets of planes. The notation <110> designates the diagonals across any face of a unit cube, and the notation {110} designates the sets of planes perpendicular to all <110> vectors. The different brackets and parentheses distinguish planes from directions and distinguish generalized planes and directions from particular planes and directions.

The mechanisms that underlie the dependence of the etch rate of an anisotropic etchant on crystal direction are not well understood. The differences in etch rate depend on temperature and on the composition of the etchant, and they also seem to be related both to the density of atomic bonds on an exposed silicon plane and to the radius of the hydrated hydroxyl ions in the etching solution.

Experience has shown, however, that exposed {111} planes are etched slowly, if at all, by the anisotropic etchants. In contrast, {110} planes are usually etched rapidly, and {100} planes are etched at an intermediate rate. Simple etched holes may have only {111} and {100} facets. Other higher-order planes, such as {221} planes, are also etched at an intermediate rate and may appear as facets of complex etched shapes.

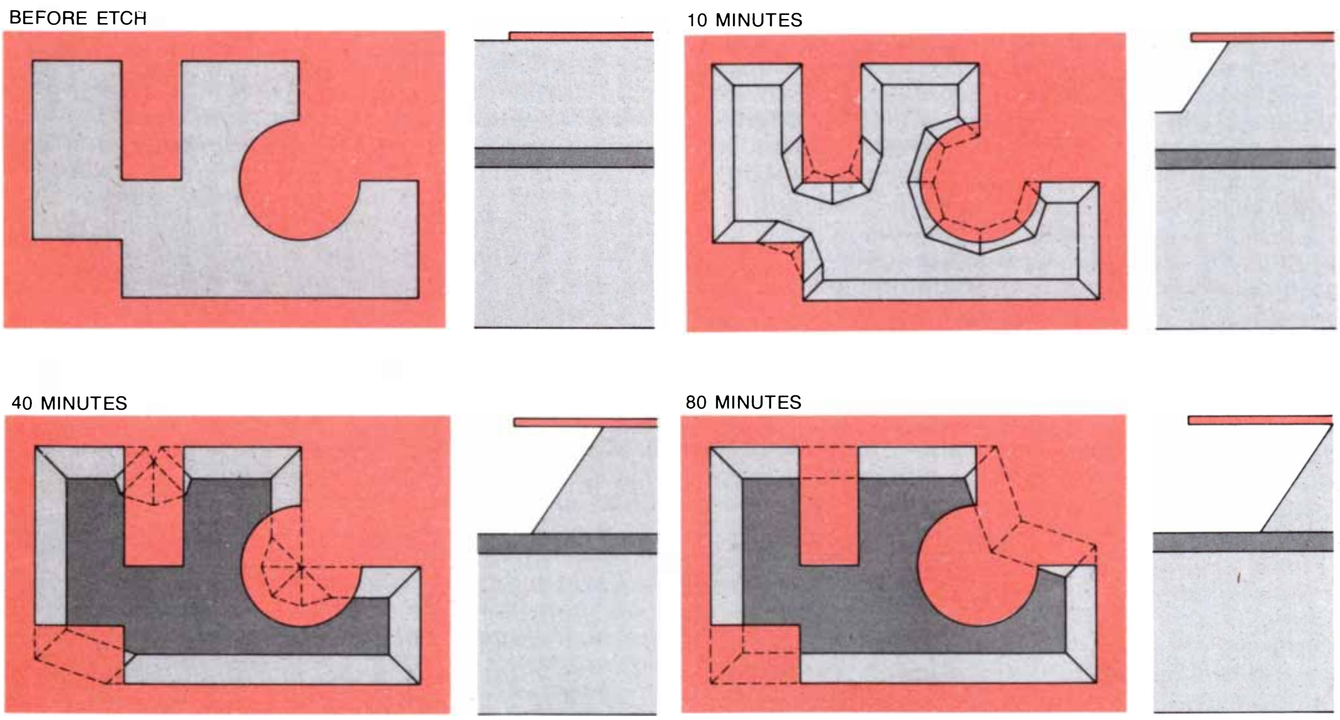
The type of hole an anisotropic etchant forms in a wafer is determined by the crystal orientation of the wafer surface as well as by the orientation dependence of the etchant. The two types of silicon wafers most useful for anisotropic etching are <100> wafers (wafers in which {100} planes are parallel to the surface and a <100> direction is perpendicular to it) and <110> wafers (in which {110} planes are parallel to the surface and a <110> direction is perpendicular to it).

The orientation, size and shape of the oxide opening on the wafer surface also play a part in determining the type of hole formed, as an example illustrates. If a square hole is opened in an oxide layer on a <100> wafer, treatment with anisotropic etchant can create an upside-down pyramidal pit. Four sets of {111} planes intersect the surface of a <100> wafer along the two perpendicular <110> directions that lie in the surface plane. During anisotropic etching the exposed {100} surface plane is etched downward at a constant rate, giving the pit a flat bottom at the start of the etch process. At the edges of the hole four inward-slanting {111} planes (at angles of 55 degrees to the surface) form the walls of the pit. As etching continues, more of the {111} planes is exposed and the area of the flat bottom shrinks. Eventually the {111} planes meet at a point and the flat bottom is gone. The etching then stops because no readily attacked planes are exposed.

The square oxide opening must be precisely aligned (within one or two degrees) with the <110> directions on the wafer surface to obtain pyramidal pits that conform exactly to the oxide mask rather than undercutting it. Most silicon wafers have a flat edge parallel to a <110> direction in the wafer. The square mask pattern is lined up with this flat edge during the photolithography step that precedes the etching.

The size of the oxide pattern determines not only the area of the pit on the wafer surface but also the depth of the pit. The larger the square oxide opening, the deeper the point at which the {111} side walls of the pit intersect. If the oxide opening is large enough, the {111} planes do not intersect within the wafer. The etched pit therefore extends all the way through the wafer, creating a small square opening on the bottom surface.

The influence of the shape of the oxide opening on the geometry of the etched pit is straightforward. If the opening in the oxide on a <100> wafer



OXIDE CANTILEVER can be formed by undercutting an oxide etching mask. Convex oxide shapes are undercut by an anisotropic etchant, although properly oriented concave shapes are not. The rectangular oxide opening shown here is a concave shape; the oxide tabs protruding into the opening are convex shapes. If the downward progress of the etchant is limited by an etch-stop layer, a prolonged anisotropic etch ultimately yields an oxide tab overhanging a shallow rectangular pit with {111} side walls and a flat bottom, namely the etch-stop layer. Cantilevers like the one formed when the oxide shape at the upper left is completely undercut have been used to make accelerometers, arrays of electrostatically controlled switches and arrays of mirrors. Undercutting makes it difficult to etch a hole into which a right-angle silicon corner projects. The silicon under the right-angle oxide shape at the lower left is rounded off during etching. A roughly right-angle silicon corner is formed under the circular oxide shape at the upper right at one point during the etch. The corner is eroded as the etch continues, however. Timed etches of patterns with circular tabs can be employed to make holes with projecting right-angle corners.

is rectangular rather than square, etching creates a long trench rather than an inverted pyramid. The walls and the ends of the trench are {111} planes, and if the etch goes to completion, the trench has a *V*-shaped cross section.

Pyramidal pits, *V*-shaped grooves and other faceted pits and holes are one kind of structural element of which micromechanical devices are composed. A second structural unit is a thin membrane of silicon. A membrane can be formed by etching a wafer for a period just short of what it would take to etch through it. It is difficult to make diaphragms of uniform thickness by this method, however, because the thickness of the diaphragm is determined by that of the wafer, and wafer thickness typically varies by 10 micrometers or more both across a single wafer and from one wafer to another.

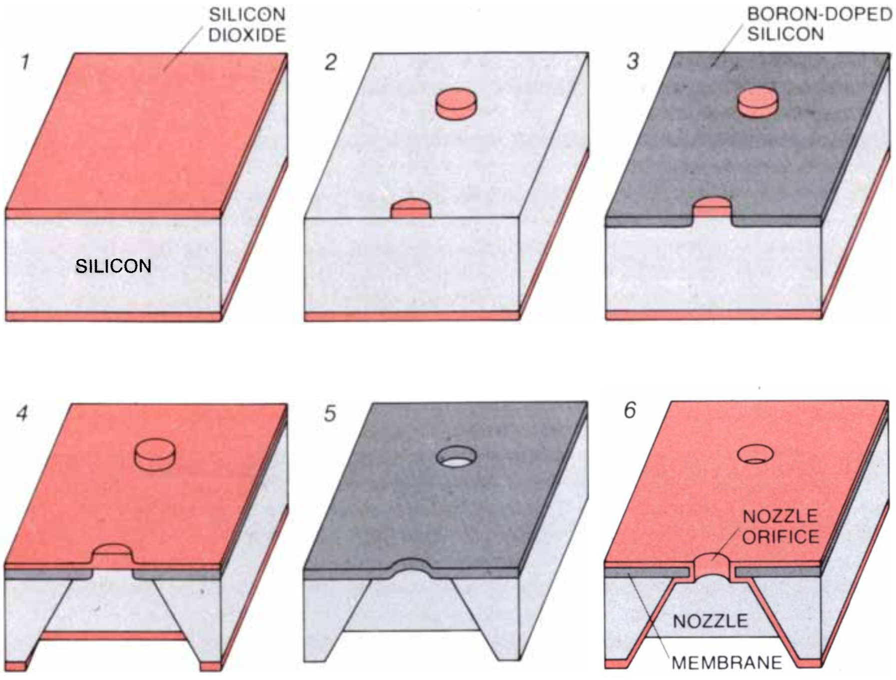
Membranes of more precisely defined thickness and shape can be created by a technique that exploits another property of anisotropic etchants. The rate at which the solutions etch a wafer depends not only on crystal orientation but also on the extent to which the silicon has been doped with impurity atoms such as boron, phosphorus, arsenic or antimony.

The etch rate of a doping-dependent etchant (a category that overlaps the categories of isotropic and anisotropic etchants) depends on the type of dopant atoms and their concentration. The isotropic etchant HNA is doping-dependent in some mixture ratios; it etches heavily doped silicon much faster than it etches lightly doped silicon. On the other hand, the anisotropic etchants potassium hydroxide and EDP etch silicon that has been heavily doped with boron much more slowly than they etch silicon lightly doped with boron.

The first step in creating a membrane by anisotropic etching is to create an etch-stop layer of silicon heavily doped with boron on the surface of the wafer. A layer of some substance rich in boron is deposited on the front surface of the wafer, which is then heated in a furnace to between 1,000 and 1,200 degrees C. At these temperatures the boron atoms diffuse into the silicon. The depth of the doped layer is controlled by the temperature and the duration of the diffusion step. If the wafer is then etched from the back with an anisotropic etchant, the etchant eats through to the doped layer and stops, creating a membrane as thick as the doped layer on the surface of the wafer.

If the designer wants to put a membrane some distance below the wafer surface, he dopes the surface in much the same way and then grows a layer of silicon over the doped layer by heating the wafer in an atmosphere of silane gas (SiH_4). The gas decomposes at high temperature and deposits silicon on the wafer surface. The new layer, which is called an epitaxial layer, assumes the crystal structure and orientation of the silicon on which it is deposited. The epitaxial layer is typically from five to 20 micrometers thick.

If the wafer is then etched from both sides in an anisotropic etchant, a membrane is created at a depth equal to the thickness of the epitaxial layer. The thickness of the membrane itself is again determined by the thickness of the boron-doped layer.



SILICON INK NOZZLE in one design for a high-speed ink-jet printer is a pyramidal pit whose bottom is a thin membrane through which an orifice of precise diameter has been etched. The diameter of the orifice is determined by the first three processing steps. The wafer is oxidized and dots of photoresist are defined on the top surface. The oxide not protected by photoresist is etched away, leaving a pattern of oxide dots. When boron is diffused into the silicon, it creates a layer of boron-doped silicon everywhere except under the dots. Then a square opening is defined in the oxide on the bottom of the wafer and is etched with an anisotropic etchant. The etchant creates a pyramidal pit with {111} side walls but stops when it comes to the doped layer. It eats through the undoped dots in the layer, however, so that when all the oxide is removed, the nozzle orifices penetrate the wafer. Finally, the wafer is reoxidized to give all surfaces a protective coating. Several arrays of nozzles are fabricated simultaneously on one wafer.

A third example of a micromechanical structure is the cantilever beam, a thin beam of silicon dioxide supported at only one end. In many applications a cantilever is suspended over a shallow pit in the silicon. The techniques for making a cantilever and a pit were developed by a group working with Kurt E. Petersen at the San Jose Research Laboratory of the International Business Machines Corporation. The techniques exploit two properties of anisotropic etchants: their doping dependence and their tendency to undercut convex shapes in an oxide mask.

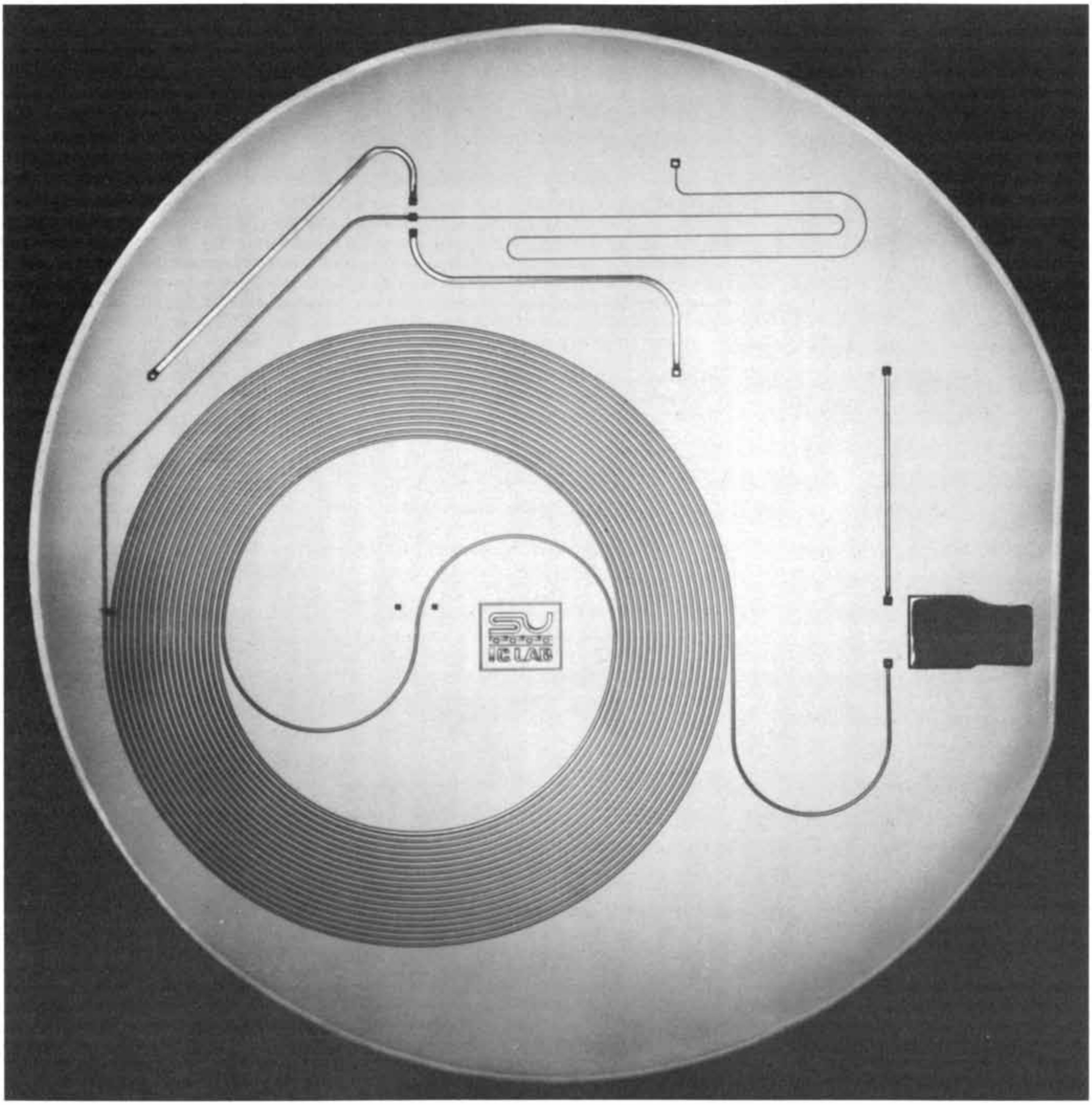
The simple geometric oxide openings discussed so far are concave oxide shapes, that is, the line between the surrounding oxide and the opening is always concave with respect to the opening. A concave oxide shape is not undercut by an anisotropic etchant if the opening is properly oriented on the wafer surface. If a finger or a tab of oxide projects from one side of such an opening, however, its boundary is convex with respect to the exposed silicon. Such convex oxide shapes are undercut by an anisotropic etchant.

The first step in making a cantilever over a shallow pit is the formation of a boron etch-stop layer. A layer of epitaxial silicon is laid down over the boron-doped stratum, oxide is grown over the epitaxial silicon and an opening is made in the oxide. The wafer is then etched in an anisotropic etchant. If an oxide tab protrudes into the oxide opening, the etchant begins to undercut it. At the attached end of the oxide tab the etchant eventually runs into a {111} plane and stops etching. Since the etch-stop layer limits the downward progress of the etch, the final result is a cantilever of oxide extending over a shallow depression in the substrate.

In the Integrated Circuits Laboratory at Stanford University work in micromachining began in 1965 when Kensall D. Wise (now at the University of Michigan) and one of us (Angell) developed a closely spaced array of electrodes for recording electric potentials in the brain. The project led to the development of a variety of brain probes. Since then projects in micromachining at Stanford have included pressure sensors, accelerometers, gas chromatographs, miniature thermometer arrays and structures for the dielectric isolation of microelectronic circuits. Micromechanical techniques and devices are also being developed at several other universities and industrial laboratories. The devices we shall describe below include some of our own and some developed by others.

Arrays of nozzles fabricated in silicon have found application in ink-jet printers for digital computers. In one type of ink-jet printer ink is forced through a linear array of nozzles, creating several parallel streams that are broken into droplets as they leave the nozzles. As droplets in each stream pass single file through a charging electrode, each droplet is either given an electric charge or allowed to remain electrically neutral. An uncharged droplet passes undeflected through an electrostatic field and leaves a dot on the paper, but a charged droplet is deflected by the field and instead of hitting the paper is intercepted by a collection gutter. A printed character is formed by selecting droplets from the vertical array of nozzles as the entire mechanism moves horizontally across the paper.

Ernest Bassous, Larry Kuhn and their colleagues at the IBM Thomas J. Watson Research Center did the pioneering work in micromechanical ink-jet nozzle arrays, developing many different nozzle systems. The simplest way to make a nozzle is to etch a pyramidal pit in the silicon. If the intersection point of the {111} planes that form the walls of the



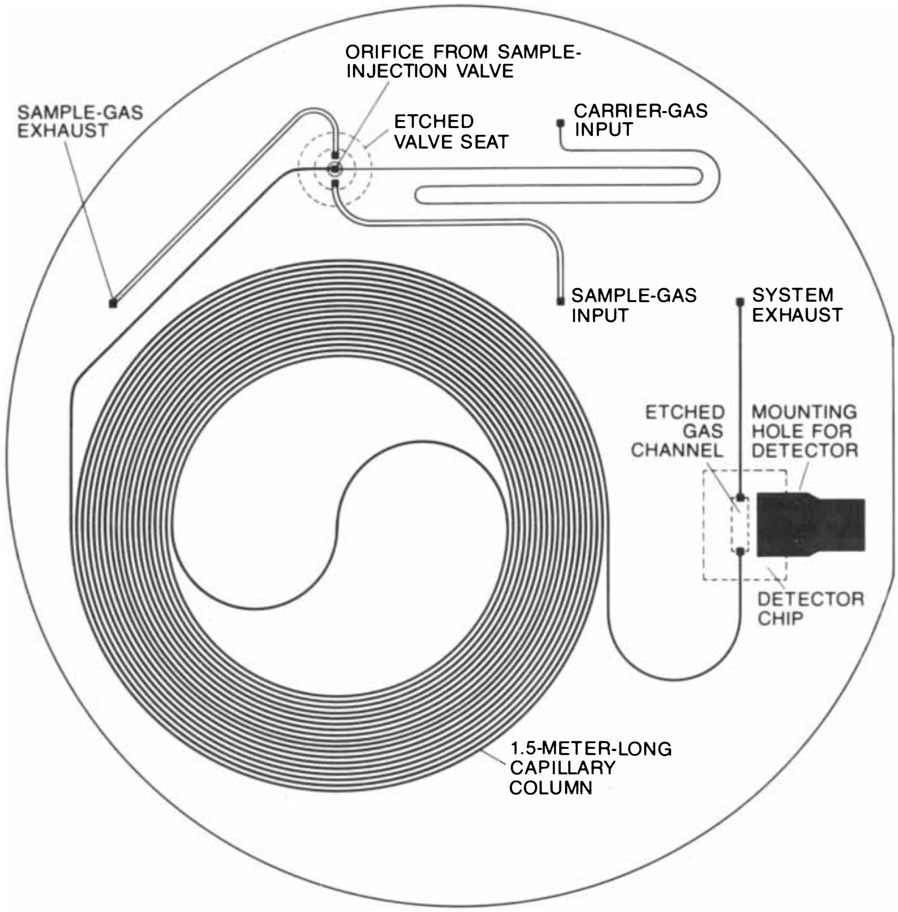
SILICON GAS CHROMATOGRAPH is a complete analytic instrument built on a wafer five centimeters across. Capillary column of the chromatograph is a spiral 1.5 meters long. The column is formed by etching a groove in a wafer and bonding the wafer to glass. This prototype was developed at Stanford University with funds provided by the National Institute for Occupational Safety and Health.

pit is below the bottom surface of the wafer, a small square opening is created. It is difficult to make nozzle orifices of uniform size with this technique, however, in part because the dimensions of the orifice depend on the thickness of the wafer; as noted above, the thickness is hard to control.

One nozzle design solves the problem by putting a membrane three micrometers thick at the bottom of each nozzle cavity. A pyramidal pit, which serves as the cavity, is etched down to the membrane, but the nozzle orifice is defined by a boron-diffusion procedure, not by the geometry of the pit. By this method Bassous and his co-workers have built arrays of eight nozzles with circular orifices exactly 20 micrometers in diameter spaced exactly .3 millimeter apart.

The first step in making the array is to define the nozzle orifices as dots of photoresist on the front surface of a wafer coated with oxide on both sides. All the oxide on the front surface is etched away except for the dots protected by photoresist. Then the wafer surface is doped with boron, which is driven into the silicon to a depth of three micrometers by heating the wafers in a diffusion furnace. Only the silicon under the oxide dots remains undoped.

Next, photolithography and an etching step are used to define a pattern of square openings in the oxide on the back surface of the wafer. The openings are centered under the undoped dots on the front surface. The wafer is etched in a doping-dependent anisotropic etchant, which eats through 99 percent of the wafer's thickness but stops when it comes to the heavily doped boron layer. It etches all the way through the undoped dots, however, creating the nozzle orifices. The diameter of each orifice, the thickness of the surrounding membrane and the placement of the orifices with respect to one another can be controlled with micrometer accuracy because they are defined not by the wafer thickness or the length of the etching step but by photolithography and the boron-diffusion step.



SEPARATION OF GASES in the silicon gas chromatograph is based on differences in the solubility of various gases in a liquid that lines the capillary column. An inert carrier gas flows continuously through the capillary-column channel. When a valve in the channel is opened, a pulse of the gas to be analyzed is fed into the column and flushed through it by the carrier gas. As the gases in the sample pass through the column they are repeatedly adsorbed and desorbed on a thin liquid lining. Each gas is identified by its retention time in the column. As each gas arrives in turn at the end of the column, it passes through a hole to a channel on the back of the wafer over which a thermal-conductivity detector is mounted. Sample gases have a lower thermal conductivity than the helium carrier gas has and cause voltage peaks in the detector output. The volume of each gas is determined from the area under the voltage peak it creates.

The most elaborate of the micromechanical devices made so far is a gas chromatograph the size of a box of kitchen matches. Two of us (Terry and Angell) built a prototype of the device in 1975. It is now being developed as a commercial product by Microsensor Technology, Inc.

Gas chromatography is an analytic technique for separating, identifying and measuring the quantity of each gas in a mixture. A sample of the mixture is injected by a valve into a long capillary column, through which it is flushed by an inert carrier gas, in most cases helium. The walls of the capillary column are lined with a thin layer of a material such as a silicone oil or a polymer in which different gases have different degrees of solubility. As the component gases pass through the column they are repeatedly adsorbed and desorbed in the lining. Because the time a component gas remains adsorbed depends on its solubility, each gas travels through the column at a different rate and emerges at a specific time. The output stream is passed over a detector that measures a property of the gas such as its thermal conductivity. The output signal generated by the detector is a series of bell-shaped peaks corresponding to the sample gases, separated by flat regions corresponding to the inert carrier gas. Many modern gas chromatographs have an associated microprocessor that identifies each gas by matching its retention time with known retention times and measures the quantity of each gas by calculating the area under the output peak.

Like a conventional gas chromatograph, the miniature one consists of a capillary column, a sample-injection valve and a thermal-conductivity detector. All the parts of the miniature gas chromatograph, however, fit on a single wafer five centimeters in diameter. The volume of the capillary column is much smaller than that of a conventional capillary column. Because a chromatograph operates properly only if the volume of the injected sample gas is much smaller than the volume of the column, it was necessary to design a miniature sample-injection valve to accompany the miniature capillary column. Because the internal volume between the valve and the column and between the column and the detector must also be minimized, the valve and column are fabricated on the same wafer and the detector is a silicon chip mounted on the wafer.

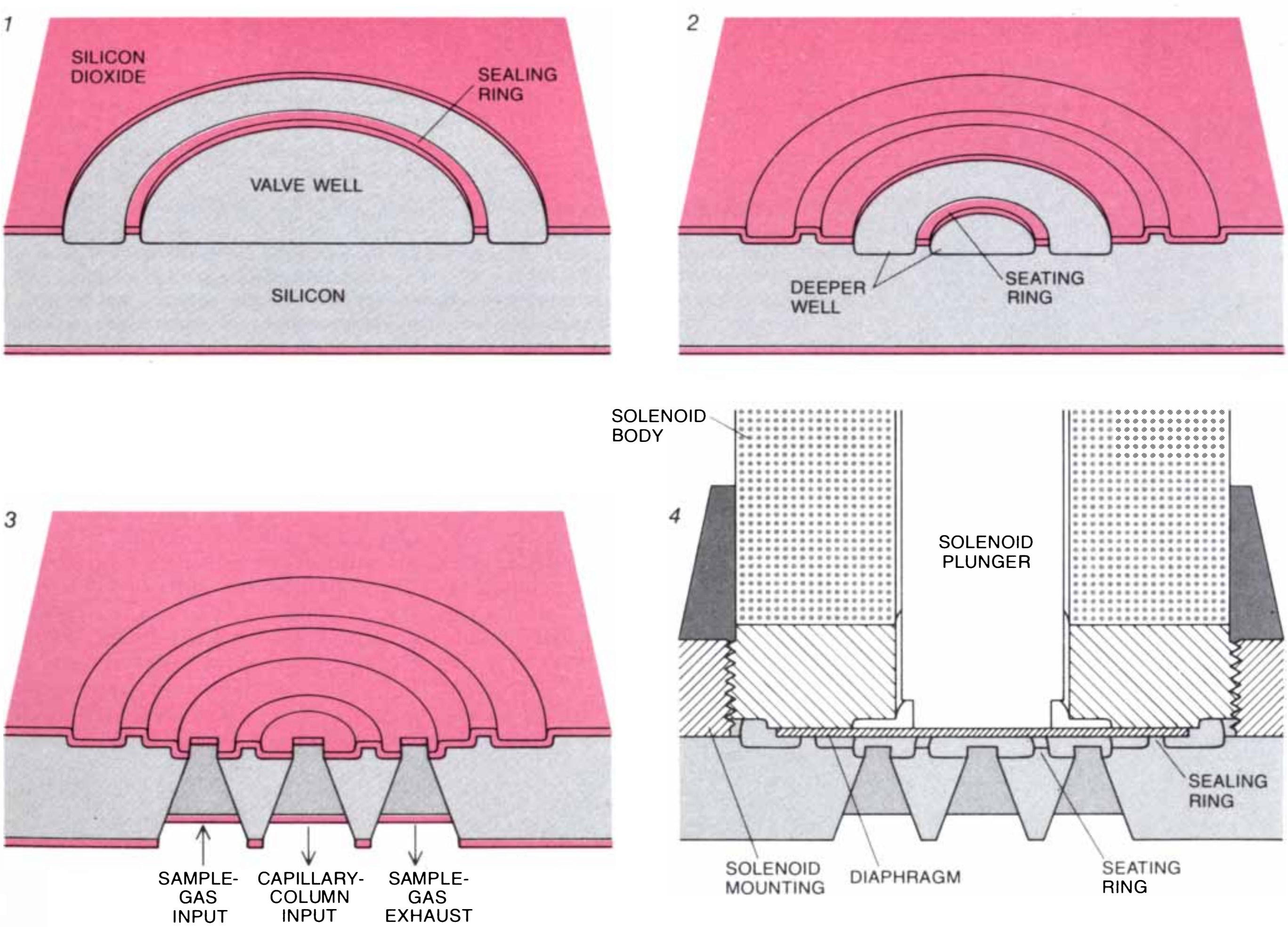
The capillary column is a groove 1.5 meters long, wound into a spiral so that it fits on the wafer. A glass plate bonded to the wafer forms the top surface of the column. At the input end of the column a hole leads to the bottom surface of the wafer. The helium carrier gas enters the column through this hole. A short distance away another hole in the column channel leads to a valve on the back surface of the wafer. The sample gas enters a separate channel through yet another hole, flows through the channel to the valve and is injected by the valve into the capillary column.

The valve seat on the back surface of the wafer consists of a silicon sealing ring, which surrounds both the input and the output orifices, and a silicon seating ring, which surrounds only the output orifice leading to the capillary column. A diaphragm of nickel and Teflon is clamped against the sealing ring. Normally a spring holds the plunger of a solenoid against the diaphragm, pushing it against the seating ring. When the solenoid is actuated, the plunger withdraws and the diaphragm relaxes, allowing gas to flow from the input orifice over the seating ring and into the output orifice. The effective dead volume of the valve is the volume of the capillary-column orifice, which is four nanoliters.

At the output end of the capillary column another etched hole leads to a gas channel etched into the bottom surface of the wafer. The chip on which the thermal-conductivity detector is built is inverted over this channel and clamped to the wafer. The detector is a thin-film metal resistor on a thermally isolating membrane of Pyrex glass in the middle of the chip. A constant electric current is passed through the resistor, and its resistance is monitored. Sample gases have a thermal conductivity lower than that of the carrier gas and remove less heat from the resistor, increasing its resistance and creating voltage peaks. The amplitude of each voltage peak is proportional to the quantity of one gas in the mixture.

The micromachining of the chromatograph wafer begins with the etching of the valve seat. The valve well, the sealing ring and the seating ring are defined by isotropic etches through concentric circular openings in the oxide layer on the back of the wafer. Then the feedthrough holes in the valve and at the ends of the capillary column and the sample-gas channels are formed by an anisotropic etch through square openings defined in the oxide on the front surface of the wafer. Finally the capillary column and the carrier-gas channels are defined by an isotropic etch through an oxide pattern on the front surface.

At this point the capillary column is a shallow spiral groove about 200 micrometers across at the surface of the wafer and 40 micrometers deep. The open spiral is made into an enclosed channel by stripping the oxide off the surface and bonding the wafer to a plate of Pyrex glass. The bonding technique, called anodic bonding, requires no intermediate liquid layer such as a glue or solder. A method of this kind is essential since



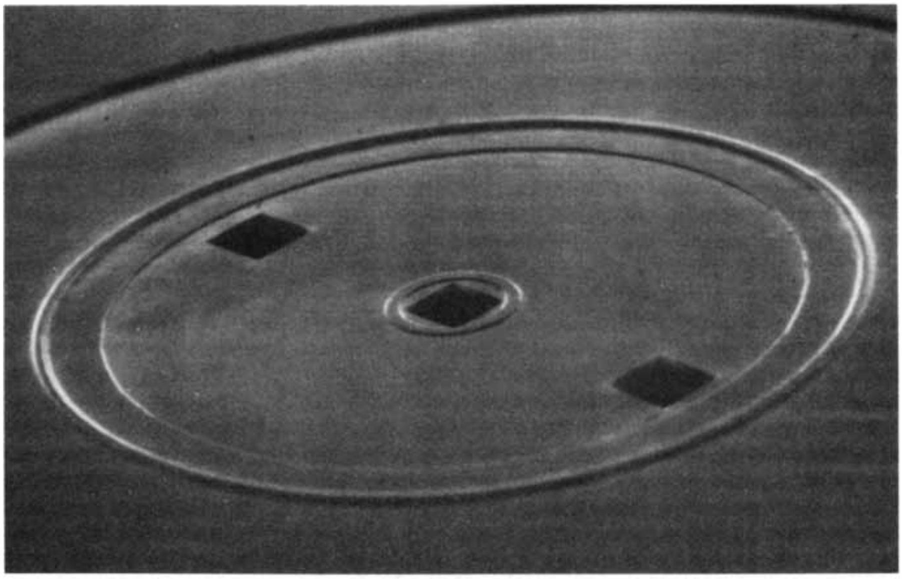
SAMPLE INJECTION VALVE for the miniature chromatograph is fabricated on the same wafer as the capillary column. Two isotropic etching steps define the valve sealing ring and the valve seat, and an anisotropic etch defines the feedthrough holes within the valve well. The first isotropic etch creates a knife-edge sealing ring of silicon raised 10 micrometers above the floor of a well. In the second step a smaller concentric well with another ring—the valve seat—is etched 10 micrometers into the floor of the first well. In the third step the feedthrough holes are formed by an anisotropic etch through square openings in the oxide on the bottom surface of the wafer. The hole within the valve seat opens to the capillary column; the other two holes open to the sample-gas channels. A metal-and-Teflon diaphragm and a solenoid plunger complete the valve. Normally the diaphragm presses against the valve seat, preventing the sample gas from flowing into the column. When the solenoid is actuated, the diaphragm relaxes, allowing pressurized sample gas to flow over the valve seat.

the shallow capillary column might be blocked by extruded glue or solder.

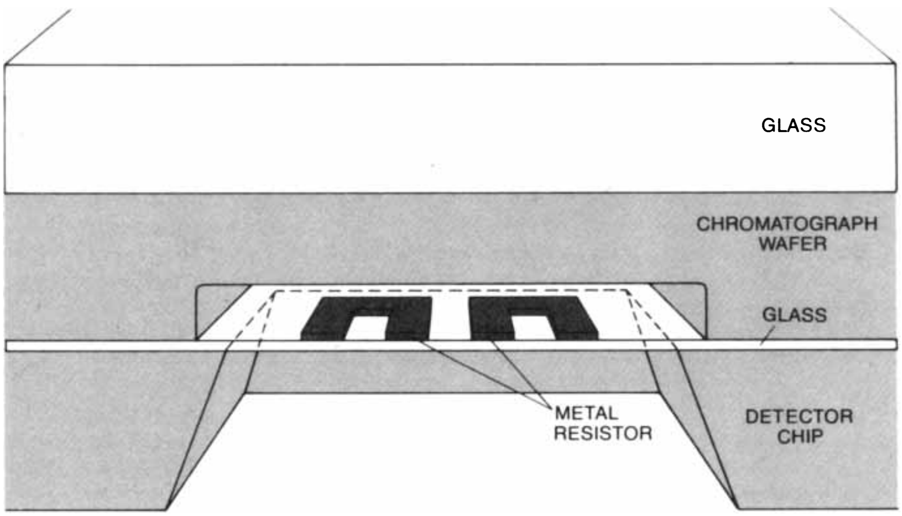
Anodic bonding, which is sometimes called electrostatic bonding or the Mallory process, was developed by George Wallis and Daniel I. Pomerantz of P. R. Mallory & Co. The glass plate and the silicon wafer are placed in contact and heated to about 400 degrees C. A large negative voltage is applied to the top surface of the glass and the silicon is electrically grounded. The glass contains a small amount of sodium; at this temperature the sodium ionizes and becomes mobile. The positive sodium ions move toward the negative electrode, leaving behind bound negative charges in the glass near the gap between the glass and the silicon. As a result the voltage across the gap is nearly equal to the applied voltage. Electrostatic attraction pulls the glass and the silicon into intimate contact and they fuse together, creating a tight seal.

The thermal-conductivity detector developed by John H. Jerman of Stanford for the gas chromatograph is a chip fabricated on a separate wafer. The wafer is oxidized and covered with a layer of Pyrex glass deposited by the technique called sputtering. Thin-film metal resistors are deposited on top of the glass. Then an anisotropic etch through square openings in the oxide on the back of the wafer removes the entire thickness of silicon, leaving a membrane of thermally isolating glass under each set of resistors. The wafer is sawed into chips, to which wires are attached, and the front surface of a chip is clamped over the gas channel at the end of the capillary column.

Because the capillary column has such a small volume, the miniature gas chromatograph requires much smaller amounts of carrier gas than a conventional chromatograph. This and the small size of the unit itself make it possible to build portable chromatography instruments. Indeed, a portable instrument consisting of five miniature chromatographs and a microcomputer is being developed by Microsensor Technology. If the five capillary columns are lined with different materials, the instrument can be made sensitive to about 100 gases. Such an instrument can monitor air quality or measure the energy content of the gas in a natural-gas pipeline.



VALVE WELL shown in this scanning electron micrograph is four millimeters in diameter and surrounds two concentric silicon rings: the valve seal (outer complete ring) and the valve seat (inner complete ring). The ring walls are 50 micrometers thick. The orifice in the center leads to the capillary column and the other two openings lead to the input and exhaust channels.



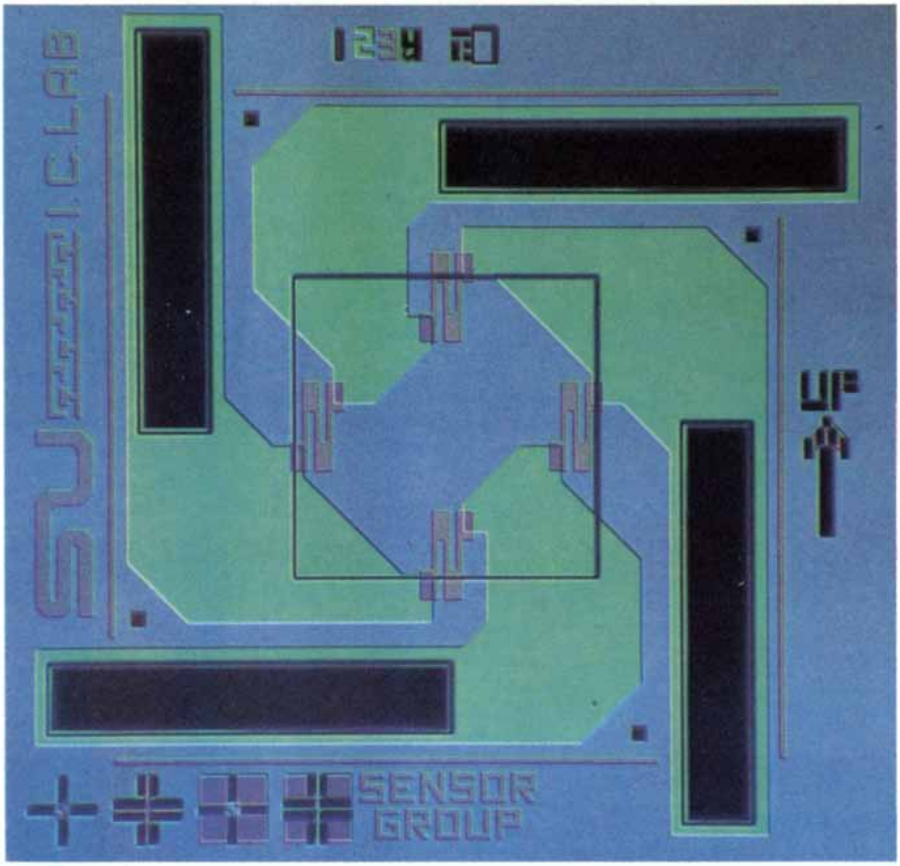
THERMAL-CONDUCTIVITY DETECTOR for the chromatograph is a resistor defined by etching a thin film of metal deposited on a wafer that has been coated with glass. After the resistor is formed the wafer is etched anisotropically from the back, leaving only a thin membrane of glass under the resistor. Because glass is a poor thermal conductor (whereas silicon is a good one), the membrane thermally isolates the resistor. A change in the thermal conductivity of the gas stream flowing over the resistor alters its temperature and hence its electrical resistance.

Micromachining has recently been used to fabricate a high-performance heat sink in a silicon wafer. The heat sink is designed to improve the performance of very-high-speed integrated circuits. In general the power dissipation of an electronic device increases along with its speed, and the heat evolved by the fastest circuits would disrupt their own operation if they were not adequately cooled. Even with special packaging schemes for high-speed circuits, the upper limit on power dissipation is about 20 watts per square centimeter of chip.

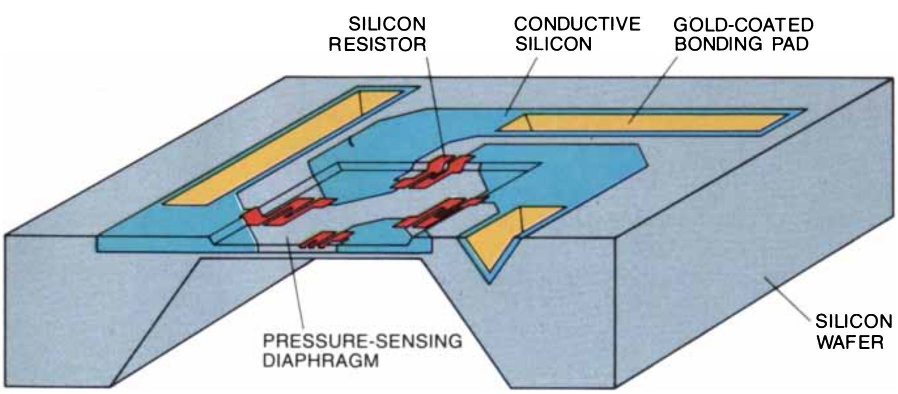
David B. Tuckerman and Fabian Pease of the Stanford Integrated Circuits Laboratory have developed a heat sink that can handle much higher power densities. The idea is to fabricate the high-speed electronic devices on the front surface of a wafer and to cut closely spaced rectangular grooves in the back of the wafer. A liquid coolant is forced through these channels. Silicon itself has very high thermal conductivity, and Tuckerman and Pease have fabricated devices in which the wall thickness, channel width and channel depth give the highest cooling efficiency.

The rectangular grooves are made in much the same way as the *V*-shaped grooves described above. An anisotropic etch through an oxide pattern consisting of narrow parallel lines creates *V* grooves if the wafer, like those discussed so far, is a <100> wafer. If the wafer is a <110> wafer, on the other hand, an anisotropic etch through the same oxide pattern can create grooves with vertical side walls and a roughly rectangular cross section. The rectangular grooves are preferable for a heat sink because they can be spaced more closely than *V* grooves and they expose more surface area to the circulating coolant. After the grooves are etched the wafer is anodically bonded to glass to create enclosed microchannels. A heat sink with etched channels 300 micrometers deep and 50 micrometers wide, separated by silicon fins 100 micrometers wide, cooled by the forced circulation of water, can handle power dissipations of more than 1,000 watts per square centimeter.

The etched heat sink can be constructed only on silicon and only on wafers of a particular orientation. Tuckerman and Pease have also investigated the possibility of creating heat sinks by sawing grooves rather than etching them. A rotary saw blade with a thin metal rim coated with diamond grit can cut channels from 50 to 200 micrometers wide on the back of a wafer. Although the dimensions of the channels cannot be controlled as precisely with sawing as they can be with etching, sawing does make it possible to create heat sinks on silicon wafers of any orientation and on wafers of other materials, such as the compound semiconductor gallium arsenide and sapphire, a substrate for some types of integrated circuit. In addition the wafers can be sawed in a crosshatch pattern, so that the circuits are supported by micropillars rather than microfins and the heat sink is less likely to be clogged by particles in the coolant.



SILICON PRESSURE SENSOR is shown in the middle of its fabrication sequence. The recessed square in the center of the photomicrograph is a thin pressure-sensing diaphragm .5 millimeter on a side. Purple serpentine elements on the diaphragm are resistors made by diffusing boron into the wafer. Green regions, which act as conductors between the resistors, have a higher boron concentration. Black rectangles near the edge of the sensor are incomplete bonding pads. They are deep *V*-shaped grooves that will be coated with gold, to which lead wires can be attached. When the diaphragm flexes, the resistance of the purple areas changes as a result of the mechanical stress, an effect called piezoresistance. The resistors are positioned so that the resistance of two of them increases and that of the other two decreases; the two-legged resistor configuration is twice as sensitive to pressure changes as a single resistor would be.



CROSS SECTION of the pressure sensor exaggerates the thickness of the diaphragm in order to show detail. The diaphragm is defined by a shallow anisotropic etch from the top of the wafer and a deep etch from the bottom. When the sensor is complete, the top surface is bonded to a glass plate, creating a sealed cavity. The bottom of the diaphragm is exposed to outside pressure. Since the pressure in the cavity is known, the output is a measure of absolute pressure.

Accelerometers and pressure sensors are two examples of integrated silicon sensors. The smallest accelerometer fabricated up to now is an oxide-cantilever device with an on-chip amplifier. The accelerometer, designed by Petersen, Anne C. Shartel and Norman F. Raley of the IBM San Jose Research Laboratory, consists of an oxide beam suspended over a shallow well formed by the boron etch-stop technique. A metal layer is deposited on the top surface of the oxide cantilever. The metal layer and the flat silicon on the bottom of the well act as two plates of a variable air-gap capacitor. A lump of gold is formed on the free end of the beam by plating. If the silicon chip is moved suddenly, the inertia of the gold weight causes the beam to flex, changing the air gap and hence the capacitance.

The output of the sensor is a voltage proportional to acceleration. The instrument has a sensitivity of two millivolts per *g* (the unit of acceleration equal to the acceleration of gravity at the earth's surface). It should be noted that the amplifier is an indispensable part of the device. The change in capacitance being detected is very small and would be swamped by larger capacitance changes in any cables attached to the sensor if the signal were not amplified on the chip. This is true in general of small capacitive sensors; signal conditioning of some kind must precede transmission.

The most successful silicon sensors at present are pressure sensors, some of which include integrated electronic devices. There are two basic types of silicon pressure sensors: piezoresistive and capacitive sensors. The electrical resistors on the thin flexible diaphragm of a piezoresistive sensor change resistance when the diaphragm flexes. In a capacitive sensor the flexible diaphragm is one plate of a variable air-gap capacitor.

In most cases the diaphragm of a pressure sensor is anisotropically etched in a <100> silicon wafer. A simple timed etch can be used to make diaphragms from 10 to 50 micrometers thick. Thinner diaphragms must be made from etch-stop layers.

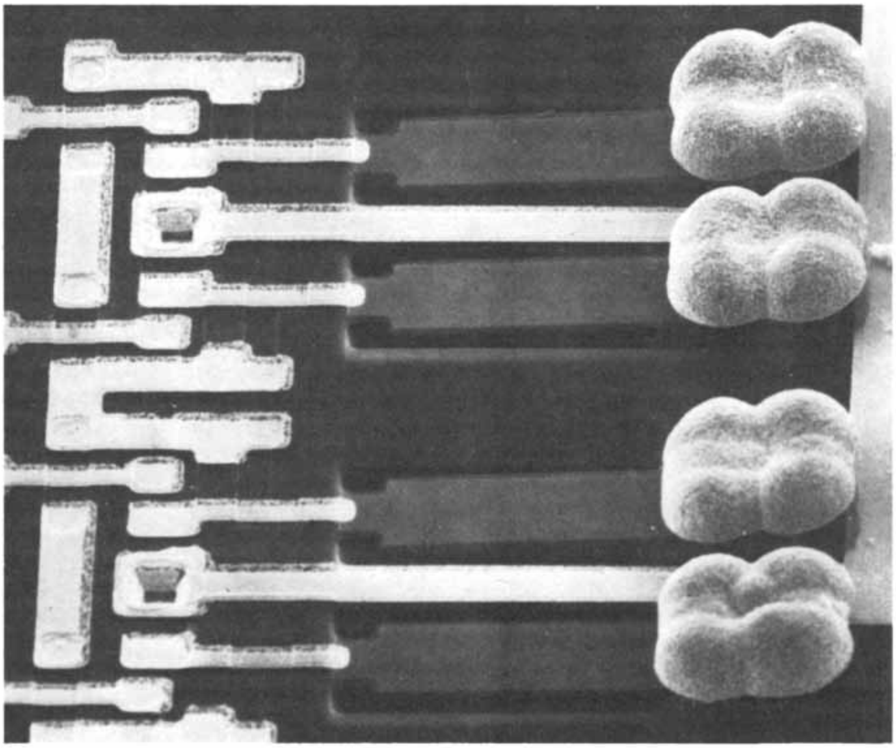
The resistors in a piezoresistive sensor are formed by doping some areas of the diaphragm. Two types of dopants are available. Elements to the right of silicon in the periodic table, such as phosphorus, have five valence electrons, whereas silicon has four. The extra electron can become a carrier of electric current. Elements to the left of silicon in the periodic table, such as boron, have three valence electrons. The absence of an electron (a "hole") can also act as a current carrier, although holes are less mobile than electrons. Current will not flow from a phosphorus-doped region to a boron-doped region because of the difference in charge carriers. Thus if the silicon from which the wafer was made was lightly doped with phosphorus, a heavier concentration of boron can be diffused into selected areas of the wafer to make resistors that are electrically isolated from the rest of the diaphragm.

The piezoresistive effect is a change in electrical resistance resulting from mechanical stress; the magnitude of the effect depends on the orientation of the resistors. By fortunate coincidence boron-doped resistors in a <100> wafer are most sensitive to stress when they are aligned with the <110> edges of an anisotropically etched cavity.

A piezoresistive pressure sensor we developed has four resistors on a square diaphragm. The resistors are near the diaphragm edges (the areas of maximum stress) and are aligned with the <110> directions (the directions along which response to stress is greatest). The resistors are joined in the two-legged configuration known as a Wheatstone bridge. When the diaphragm flexes, the resistance of two resistors increases and that of the other two decreases by an equal amount. As a result the bridge is twice as sensitive to pressure changes as a single resistor would be.

The first pressure-sensor chip to include active circuitry (that is, circuitry including transistors that amplify the signal) was designed by John M. Borky and Kensall Wise of the University of Michigan. The circuitry converts the output voltage of a Wheatstone bridge into a frequency-modulated (FM) signal; in other words, changes in pressure are translated into changes in frequency. If a different range of frequencies is assigned to each of several sensors, they can share a transmission line in the same way different FM radio stations share the radio-frequency band.

Capacitive pressure sensors are inherently more sensitive to pressure changes than piezoresistive sensors, but as we pointed out above, the signal from a capacitive sensor can easily be lost in noise. Signal-conditioning circuitry improves the performance of a piezoresistive sensor, but it is crucial to that



CAPACITIVE ACCELEROMETERS are cantilever beams of silicon dioxide overhanging shallow wells in the silicon. Four accelerometers are shown in this scanning electron micrograph. Bumps of gold at the right end of each cantilever increase the inertia of the beam and therefore the sensitivity of the device. A metal layer on top of the beam, an electrically conductive boron etch-stop layer at the bottom of the well and the air gap between them form a variable capacitor. Under acceleration the beam flexes, changing the capacitance. The circuitry at the left amplifies the signal; the sensor output is a voltage proportional to acceleration.

of a capacitive sensor. Craig S. Sander, James W. Knutti and James D. Meindl of the Stanford Integrated Circuits Laboratory designed the first capacitive pressure sensor with on-chip active circuitry. The diaphragm of this sensor, which is recessed about five micrometers into the silicon wafer, serves as one plate of a capacitor. The other plate is a metal film on a sheet of glass that is anodically bonded to the wafer.

A fixed quantity of gas is sealed in the cavity between the diaphragm and the metal-coated glass plate. The other side of the diaphragm is exposed to the pressure to be measured. When the pressure changes and the diaphragm flexes, the thickness of the insulating air gap between the metal and the silicon changes. The active circuitry on the chip monitors the resulting changes in capacitance, putting out a series of voltage pulses. The time between two pulses is proportional to the measured pressure. This pulse-time-coded signal, like the FM signal from the integrated piezoresistive pressure sensor, is less easily obscured by noise than a simple uncoded voltage would be.

Integrated sensors can meet the requirements of a wider variety of applications than discrete sensors because the signals are relatively immune to noise and leakage. Thus integrated sensors can be put in harsh environments, such as under the hood of an automobile, or in situations where the sensing device must be physically separated from the signal-processing electronics, as in biomedical monitoring. Many of the sensors now under development are designed for such difficult applications. For example, two of us (Barth and Angell) are working on sensors to be implanted in living tissue.

Also under development are chips with several types of sensors, such as a temperature and a pressure sensor, several copies of the same sensor (for reliability) and chips that include some means of sensor recalibration. In the future many integrated sensors will be able to convert analogue signals into digital signals so that they can be connected directly to digital electronics. In general the biggest challenge at present is to determine how much circuitry it is feasible and cost-effective to put on a sensor chip.

At the same time that a great deal of attention is being given to sensors, strictly mechanical silicon devices continue to be developed. Silicon etching techniques are still being refined, and new micromechanical structures will emerge from the laboratories over the next several years. Silicon mechanical devices may one day be nearly as commonplace as silicon electronic devices are now.

The Structure of Quarks and Leptons

They have been considered the elementary particles of matter, but instead they may consist of still smaller entities confined within a volume less than a thousandth the size of a proton

by Haim Harari

In the past 100 years the search for the ultimate constituents of matter has penetrated four layers of structure. All matter has been shown to consist of atoms. The atom itself has been found to have a dense nucleus surrounded by a cloud of electrons. The nucleus in turn has been broken down into its component protons and neutrons. More recently it has become apparent that the proton and the neutron are also composite particles; they are made up of the smaller entities called quarks. What comes next? It is entirely possible that the progression of orbs within orbs has at last reached an end and that quarks cannot be more finely divided. The leptons, the class of particles that includes the electron, could also be elementary and indivisible. Some physicists, however, are not at all sure the innermost kernel of matter has been exposed. They have begun to wonder whether the quarks and leptons too might not have some internal composition.

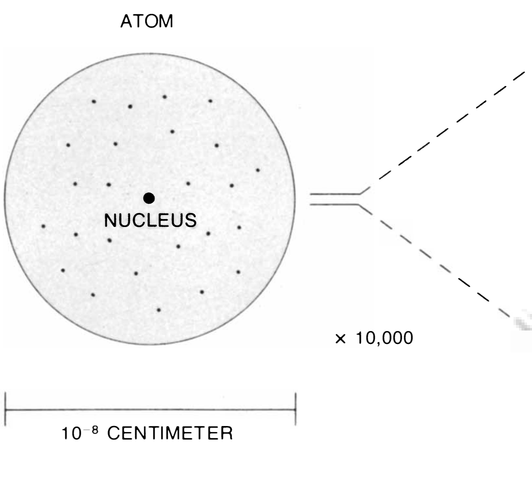
The main impetus for considering still another layer of structure is the conviction (or perhaps prejudice) that there should be only a few fundamental building blocks of matter. Economy of means has long been a guiding principle of physics, and it has served well up to now. The list of the basic constituents of matter first grew implausibly long toward the end of the 19th century, when the number of chemical elements, and hence the number of species of atoms, was approaching 100. The resolution of atomic structure solved the problem, and in about 1935 the number of elementary particles stood at four: the proton, the neutron, the electron and the neutrino. This parsimonious view of the world was spoiled in the 1950's and 1960's; it turned out that the proton and the neutron are representatives of a very large family of particles, the family now called hadrons. By the mid-1960's the number of fundamental forms of matter was again roughly 100. This time it was the quark model that brought relief. In the initial formulation of the model all hadrons could be explained as combinations of just three kinds of quarks.

Now it is the quarks and leptons themselves whose proliferation is beginning to stir interest in the possibility of a simpler scheme. Whereas the original model had three quarks, there are now thought to be at least 18, as well as six leptons and a dozen other particles that act as carriers of forces. Three dozen basic units of matter are too many for the taste of some physicists, and there is no assurance that more quarks and leptons will not be discovered. Postulating a still deeper level of organization is perhaps the most straightforward way to reduce the roster. All the quarks and leptons would then be composite objects, just as atoms and hadrons are, and would owe their variety to the number of ways a few smaller constituents can be brought together. The currently observed diversity of nature would be not intrinsic but combinatorial.

It should be emphasized that as yet there is no evidence quarks and leptons have an internal structure of any kind. In the case of the leptons, experiments have probed to within 10^{-16} centimeter and found nothing to contradict the assumption that leptons are pointlike and structureless. As for the quarks, it has not been possible to examine a quark in isolation, much less to discern any possible internal features. Even as a strictly theoretical conception, the subparticle idea has run into difficulty: no one has been able to devise a consistent description of how the subparticles might move inside a quark or a lepton and how they might interact with one another. They would have to be almost unimaginably small: if an atom were magnified to the size of the earth, its innermost constituents could be no larger than a grapefruit. Nevertheless, models of quark and lepton substructure make a powerful appeal to the aesthetic sense and to the imagination: they suggest a way of building a complex world out of a few simple parts.

Any theory of the elementary particles of matter must also take into account the forces that act between them and the laws of nature that govern the forces. Little would be gained in simplifying the spectrum of particles if the number of forces and laws were thereby increased. As it happens, there has been a subtle interplay between the list of particles and the list of forces throughout the history of physics.

In about 1800 four forces were thought to be fundamental: gravitation, electricity, magnetism and the short-range force between molecules that is responsible for the cohesion of matter. A series of remarkable experimental and theoretical discoveries then led to the recognition that electricity and magnetism are actually two manifestations of the same basic force, which was soon given the name electromagnetism. The discovery of atomic structure brought a further revision. Although an atom is



HIERARCHY OF PARTICLES in the structure of matter currently has four levels. All matter is made up of atoms; the atom con-electrically neutral overall, its constituents are charged, and the short-range molecular force came to be understood as a complicated residual effect of electromagnetic interactions of positive nuclei and negative electrons. When two neutral atoms are far apart, there are practically no electromagnetic forces between them. When they are near each other, however, the charged constituents of one atom are able to "see" and influence the inner charges of the other, leading to various short-range attractions and repulsions.

As a result of these developments physics was left with only two basic forces. The unification of electricity and magnetism had reduced the number by one, and the molecular interaction had been demoted from the rank of a fundamental force to that of a derivative one. The two remaining fundamental forces, gravitation and electromagnetism, were both long-range. The exploration of nuclear structure, however, soon introduced two new short-range forces. The strong force binds protons and neutrons together in the nucleus, and the weak force mediates certain transformations of one particle into another, as in the beta decay of a radioactive nucleus. Thus there were again four forces.

The development of the quark model and the accompanying theory of quark interactions was the next occasion for revising the list of forces. The quarks in a proton or a neutron are thought to be held together by a new long-range fundamental force called the color force, which acts on the quarks because they bear a new kind of charge called color. (Neither the force nor the charge has any relation to ordinary colors.) Just as an atom is made up of electrically charged constituents but is itself neutral, so a proton or a neutron is made up of colored quarks but is itself colorless. When two colorless protons are far apart, there are essentially no color forces between them, but when they are near, the colored quarks in one proton "see" the color charges in the other proton. The short-range attractions and repulsions that result have been identified with the effects of the strong force. In other words, just as the short-range molecular force became a residue of the long-range electromagnetic force, so the short-range strong force has become a residue of the long-range color force.

One more chapter can be added to this abbreviated history of the forces of nature. A deep and beautiful connection has been found between electromagnetism and the weak force, bringing them almost to the point of full unification. They are clearly related, but the connection is not quite as close as it is in the case of electricity and magnetism, and so they must still be counted as separate forces. Therefore the current list of fundamental forces still has four entries: the long-range gravitational, electromagnetic and color forces and the short-range weak force. Within the limits of present knowledge all natural phenomena can be understood through these forces and their residual effects.

The evolution of ideas about particles and that of ideas about forces are clearly interdependent. As new basic particles are found, old ones turn out to be composite objects. As new forces are discovered, old ones are unified or reduced to residual status. The lists of particles and forces are revised from time to time as matter is explored at smaller scale and as theoretical understanding progresses. Any change in one list inevitably leads to a modification of the other. The recent speculations about quark and lepton structure are no exception; they too call for changes in the complement of forces. Whether the changes represent a simplification remains to be seen.

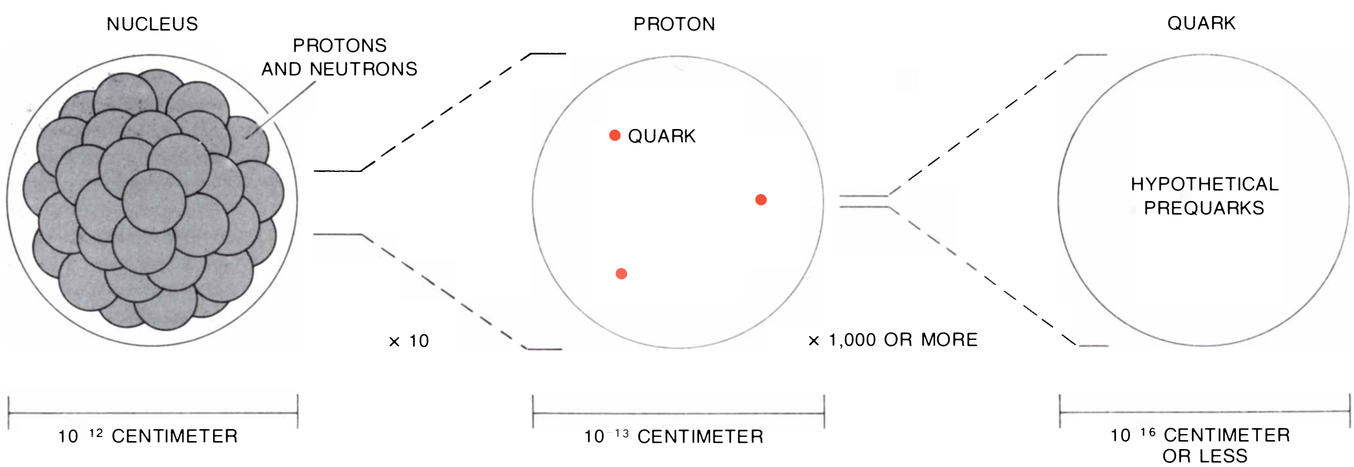
Of the four established fundamental forces, gravitation must be put in a category apart. It is too feeble even to be detected in the interactions of individual particles, and it is not understood in terms of microscopic events. For the other three forces successful theories have been developed and are now widely accepted. The three theories are distinct, but they are consistent with one another; taken together they constitute a comprehensive model of elementary particles and their interactions, which I shall refer to as the standard model.

In the standard model the indivisible constituents of matter are the quarks and the leptons. It is convenient to discuss the leptons first. There are six of them: the electron and its companion the electron-type neutrino, the muon and the muon-type neutrino and the tau and the tau-type neutrino. The electron, the muon and the tau have an electric charge of -1; the three neutrinos are electrically neutral.

There are also six basic kinds of quark, which have been given the names up, down, charmed, strange, top and bottom, or *u, d, c, s, t* and *b.* (The top quark has not yet been detected experimentally, and neither has the tau-type neutrino, but few theorists doubt their existence.) The *u, c* and *t* quarks have an electric charge of $+2/3$, the *d, s* and *b* quarks a charge of $-1/3$. In addition each quark type has three possible colors, which I shall designate red, yellow and blue. Thus if each colored quark is counted as a separate particle, there are 18 quark varieties altogether. Note that each quark carries both color and electric charge, but none of the leptons are colored.

For each particle in this scheme there is an antiparticle with the same mass but with opposite values of electric charge and color. The antiparticle of the electron is the positron, which has a charge of $+1$. The antiparticle of a red *u* quark, with a charge of $+2/3$, is an antired $\bar{u}$ antiquark, with a charge of $-2/3$.

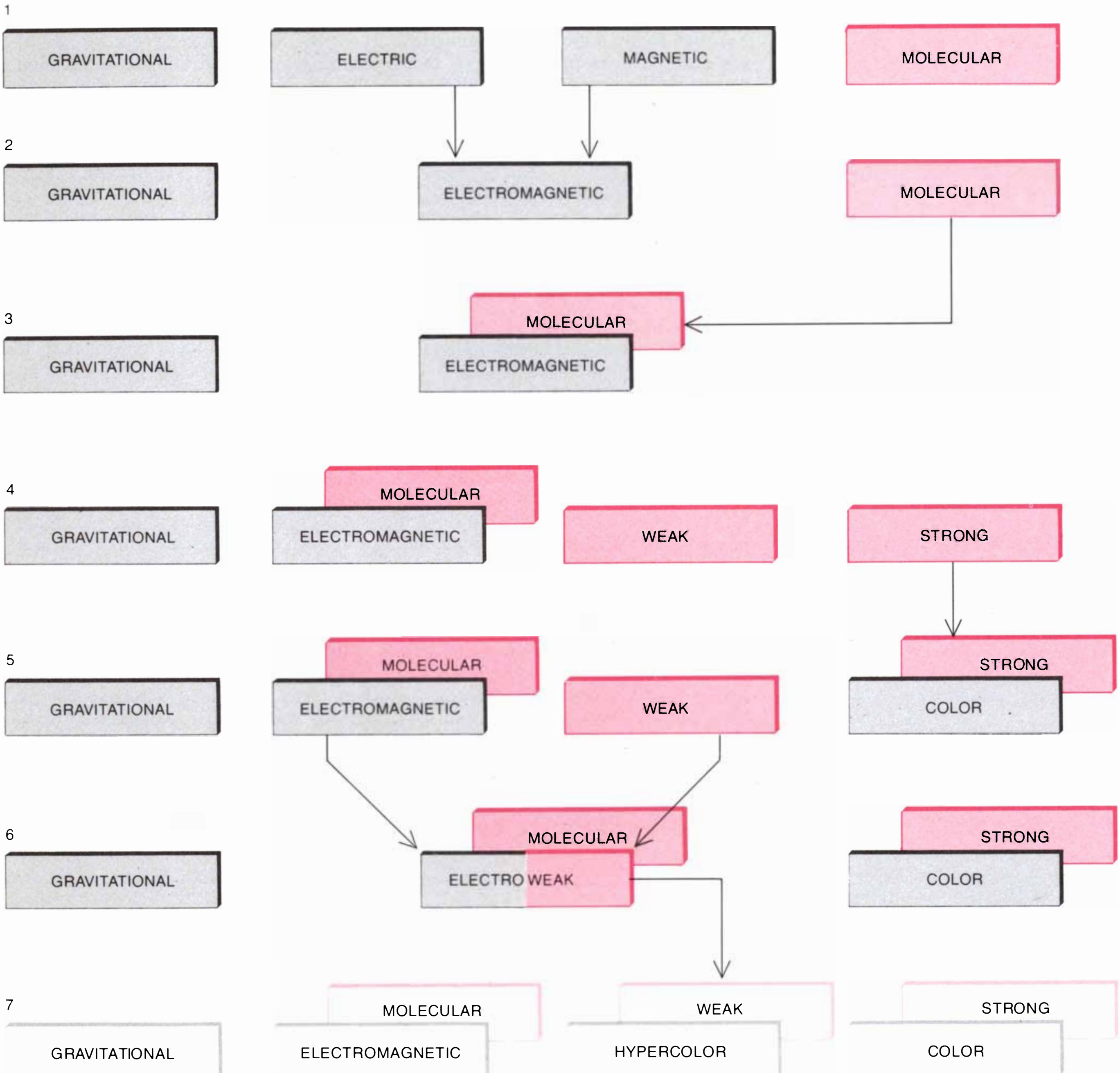
The color property of the quarks is analogous in many ways to electric charge, but because there are three pos-



sists of a nucleus surrounded by electrons; the nucleus is composed of protons and neutrons; each proton and neutron is thought to be composed of three quarks. Recent speculations might add a fifth level: the quark might be a composite of hypothetical finer constituents, which can be generically called prequarks. The leptons, the class of particles that includes the electron, could also consist of prequarks.

sible colors it is appreciably more complicated. Electrically charged particles can be brought together to form an electrically neutral system in only one way: by combining equal quantities of positive and negative charge. A colorless composite particle can be formed out of colored quarks in much the same way, namely by combining a colored quark and an anticolored antiquark. In the case of color, however, there is a second way to form a neutral state: any composite system with equal quantities of all three colors or of all three anticolors is also colorless. For this reason a proton consisting of one red quark, one yellow quark and one blue quark has no net color.

One further property of the quarks and leptons should be mentioned: each particle has a spin, or intrinsic angular momentum, equal to one-half the basic quantum-mechanical unit of angular momentum. When a particle with a spin of 1/2 moves along a straight line, its intrinsic rotation can be either clockwise or counterclockwise when the particle is viewed along the direction of motion. If the spin is clockwise, the particle is said to be right-handed, because when the fingers of the right hand curl in the same direction as the spin, the thumb indicates the direction of motion. For a par-



FUNDAMENTAL FORCES OF NATURE can be classified in a scheme that has evolved together with the list of elementary particles. Here long-range forces are shown in gray and short-range ones in color. Early in the 19th century three long-range forces were thought to be fundamental: the gravitational, the electric and the magnetic. One short-range force, the molecular force responsible for the cohesion of matter, also had fundamental status. James Clerk Maxwell unified electricity and magnetism, and with the discovery of atomic structure it became apparent that the molecular force is not fundamental but instead is a residual effect of electromagnetism. The discovery of the atomic nucleus introduced two short-range forces, the weak and the strong. In the quark model, however, the strong force becomes a residue of a new long-range force called the color force. Furthermore, a deep relation has been found between the weak force and electromagnetism, so that they can be considered partially unified. The sixth row of the chart represents the forces of nature as they are now understood in the "standard model" of elementary-particle physics. A successful model of quark and lepton structure could bring a further revision. In some models, for example, there is a new long-range force called hypercolor, and the weak force is a residue of it. In models of this kind all the fundamental forces of nature are long-range ones. Such models, however, are still highly speculative.

ticle with the opposite sense of spin a left-hand rule describes the motion, and so the particle is said to be left-handed.

In the standard model the three forces that act on the quarks and leptons are described by essentially the same mathematical structure. It is known as a gauge-invariant field theory or simply a gauge theory. Each force is transmitted from one particle to another by carrier fields, which in turn are embodied in carrier particles, or gauge bosons.

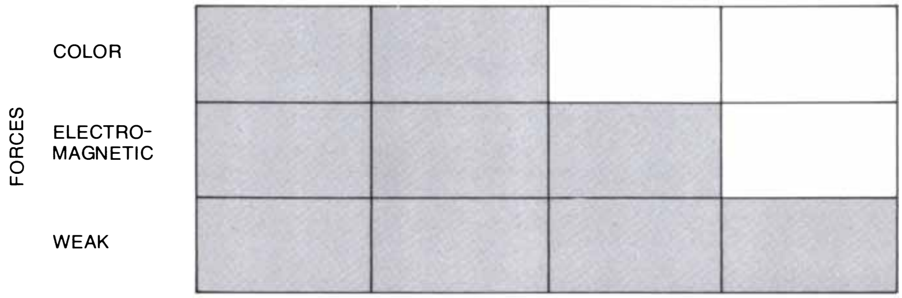
The gauge theory of the electromagnetic force, called quantum electrodynamics or QED, is the earliest and simplest of the three theories. It was devised in the 1940's by Richard P. Feynman, Julian S. Schwinger and Sin-Itiro Tomonaga. QED describes the interactions of electrically charged particles, most notably the electron and the positron. There is one kind of gauge boson to mediate the interactions; it is the photon, the familiar quantum of electromagnetic radiation, and it is massless and has no electric charge of its own. QED is probably the most accurately tested theory in physics. For example, it correctly predicts the magnetic moment of the electron to at least 10 significant digits.

The theory of the color force was formulated by analogy to QED and is called quantum chromodynamics or QCD. It was developed over a period of almost two decades through the efforts of many theoretical physicists. In QCD particles interact by virtue of their color rather than their electric charge. The gauge bosons of QCD, which are responsible for binding quarks inside a hadron, are called gluons. Like the photon, the gluons are massless, but whereas there is just one kind of photon, there are eight species of gluons. A further difference between the photon and the gluons turns out to be even more important. Although the photon is the intermediary of the electromagnetic force, it has no electric charge and hence gives rise to no electromagnetic forces of its own (or at least none of significant magnitude). The gluons, in contrast, are not colorless. They transmit the color force between quarks but they also have color of their own and respond to the color force. This reflexiveness, whereby the carrier of the force acts on itself, makes a complete mathematical analysis of the color force exceedingly difficult.

One peculiarity that seems to be inherent in QCD is the phenomenon of color confinement. It is thought that the color force somehow traps colored objects (such as quarks and gluons) inside composite objects that are invariably colorless (such as protons and neutrons). The colored particles can never escape (although they can form new colorless combinations). It is because of color confinement, physicists suppose, that a quark or a gluon has never been seen in

		QUARKS		LEPTONS	
GENERATIONS	FIRST	u (UP)	d (DOWN)	e (ELECTRON)	ν_e (ELECTRON-TYPE NEUTRINO)
	SECOND	c (CHARMED)	s (STRANGE)	μ (MUON)	ν_μ (MUON-TYPE NEUTRINO)
	THIRD	t (TOP)	b (BOTTOM)	τ (TAU)	ν_τ (TAU-TYPE NEUTRINO)

CHARGES	ELECTRIC	+2/3	−1/3	−1	0
	COLOR	RED YELLOW BLUE	RED YELLOW BLUE	COLORLESS	COLORLESS



STANDARD MODEL of elementary particles includes three "generations" of quarks and leptons, although all ordinary matter can be constructed out of the particles of the first generation alone. The quarks are distinguished by fractional values of electric charge and by a property that is fancifully called color: each quark type comes in red, yellow and blue versions. The leptons have integer units of electric charge and are colorless. The two classes of particles also differ in their response to the various forces. Only the quarks are subject to the color force, and as a result they may be permanently confined inside composite particles such as the proton.

isolation. I must stress that although the idea of color confinement is now widely accepted, it has not been proved to follow from QCD. There may still be surprises in store.

The weak force is somewhat different from the other two, but it can nonetheless be described by a gauge theory of the same general kind. The theory was worked out, and the important connection between the weak force and electromagnetism was established, in the 1960's and the early 1970's by a large number of investigators. Notable contributions were made (in chronological order) by Sheldon Lee Glashow of Harvard University, Steven Weinberg of the University of Texas at Austin, Abdus Salam of the International Centre for Theoretical Physics in Trieste and Gerard 't Hooft of the University of Utrecht.

Curiously, the charges on which the weak force acts are associated with the handedness of a particle. Among both quarks and leptons left-handed particles and right-handed antiparticles have a weak charge, but right-handed particles and left-handed antiparticles are neutral with respect to the weak force. What is odder still, the weak charge is not conserved in nature: a unit of charge can be created out of nothing or can disappear into the vacuum. In contrast, the net quantity of electric charge in an isolated system of particles can never be altered, and neither can the net color. The weak force is also distinguished by its exceedingly short range; its effects extend only to a distance of about 10^{-16} centimeter, or roughly a thousandth of the diameter of a proton.

In the gauge theory of the weak force both the failure of the weak charge to be conserved and the short range of the force are attributed to a mechanism called spontaneous symmetry breaking, which I shall discuss in greater detail below. For now it is sufficient to note that the symmetry-breaking mechanism implies that the weak charge, and the associated handedness of particles, should be conserved at extremely high energy, where a particle's mass is a negligible fraction of its kinetic energy.

Spontaneous symmetry breaking also requires that the gauge bosons of the weak force be massive particles; indeed, they have masses approximately 100 times the mass of the proton. In the standard model there are three such bosons: two of them, designated W^+ and W^-, carry electric charge as well as weak charge; the third, designated Z^0, is electrically neutral. The large mass of the weak bosons accounts for the short range of the force. According to the uncertainty principle of quantum mechanics, the range of a force is inversely proportional to the mass of the particle that transmits it. Thus electromagnetism and the color force, being carried by massless gauge bosons, are effectively infinite in range, whereas the weak force has an exceedingly small sphere of influence. Spontaneous symmetry breaking has still another consequence: it predicts the existence of at least one additional massive particle, separate from the weak bosons. It is called the Higgs particle after Peter Higgs of the University of Edinburgh, who made an important contribution to the theory of spontaneous symmetry breaking.

In the past 10 years the successes of the standard model have given physicists a good deal of self-confidence. All known forms of matter can be constructed out of the 18 colored quarks and the six leptons of the model. All observed interactions of matter can be explained as exchanges of the 12 gauge bosons included in the model: the photon, the eight gluons and the three weak bosons. The model seems to be internally consistent; no one part is in conflict with any other part, and all measurable quantities are predicted to have a plausible, finite value. Internal consistency is not a trivial achievement in a conceptual system of such wide scope. So far the model is also consistent with all experimental results, that is to say, no clear prediction of the model has yet been contradicted by experiment. To be sure, there are some important predictions that have not yet been fully verified; most notably, the tau-type neutrino, the top quark, the weak bosons and the Higgs particle must be found. The first direct evidence of W bosons was recently reported by a group of experimenters at CERN, the European Laboratory for Particle Physics in Geneva. In the next several years new particle accelerators and more sensitive detecting apparatus will test the remaining predictions of the model. Most physicists are quite certain they will be confirmed.

If the standard model has proved so successful, why would anyone consider more elaborate theories? The primary motivation is not a suspicion that the standard model is wrong but rather a feeling that it is less than fully satisfying. Even if the model gives correct answers for all the questions it addresses, many questions are left unanswered and many regularities in nature remain coincidental or arbitrary. In short, the model itself stands in need of explanation.

The strongest hint of some organizing principle beyond the standard model is the proliferation of elementary particles. The known properties of matter are not so numerous or diverse that 24 particles are needed to represent them all. Indeed, there seems to be a great deal of repetition in the spectrum of quarks and leptons. There are three leptons with an electric charge of −1, three neutral leptons, three quarks with a charge of +2/3 and three quarks with a charge of −1/3. Everything is triplicated, and for no apparent reason. A world constructed by choosing one particle from each of the four groups would seem to have all the necessary variety.

As it turns out, all ordinary matter can indeed be formed from a subset that includes just the *u* quark, the *d* quark, the electron and the electron-type neutrino. These four particles and their antiparticles make up the "first generation" of quarks and leptons. The remaining quarks and leptons merely repeat the same pattern in two additional generations without seeming to add anything new. Corresponding particles in different generations are identical in all respects except one: they have different masses. The *d, s* and *b* quarks, for example, respond in precisely the same way to the electromagnetic, color and weak forces. For some unknown reason, however, the *s* quark is roughly 20 times as heavy as the *d* quark, and the *b* quark is approximately 600 times as heavy as the *d*. The mass ratios of the other quarks and of the charged leptons are likewise large and unexplained. (The masses of the neutrinos are too small to have been measured; it is not yet known whether the neutrinos are merely very light or are entirely massless.)

The presence of three generations of quarks and leptons begs for an explanation. Why does nature repeat itself? The pattern of particle masses is also mysterious. In the standard model the masses are determined by approximately 20 "free" parameters that can be assigned any values the theorist chooses; in practice the values are generally based on experimental findings. Is it possible the 20 parameters are all unrelated? Are they fundamental constants of nature with the same status as the velocity of light or the electric charge of the electron? Probably not.

A further tantalizing regularity can be perceived in the electric charges of the quarks and leptons: they are all related by simple ratios and are all integer multiples of one-third the electron charge. The standard model supplies no reason; in principle the charge ratios could have any values. It can be deduced from observation that the ratios of one-third and two-thirds that define the quark charges are not approximations. The proton consists of two *u* quarks and a *d* quark, with charges of 2/3 + 2/3 − 1/3, or +1. If these values were not exact and the quarks instead had charges of, say, +.617 and −.383, the magnitude of the proton's charge would not be exactly equal to that of the electron's, and ordinary atoms would not be electrically neutral. Since atoms can be brought together in enormous numbers, even a slight departure from neutrality could be readily detected.

If the particles and antiparticles that make up a single generation are arranged according to their charge, it is found that every value from −1 to +1 in intervals of one-third is occupied by one particle (or, in the case of zero charge, by two particles, namely the neutrino and the antineutrino). The pattern formed raises still more questions. Why has nature favored these values of electric charge but no others, such as +4/3 or −5/3? It is apparent that all particles with integral charge are colorless and all those with fractional charge are colored. Is there some relation between the electric charge of a particle and its color or between the quarks and the leptons? The standard model implies no such relations, but they seem to exist.

Another motivation for looking beyond the standard model is the continuing desire to unify the fundamental forces, or at least to find some relation

ELECTRIC CHARGE

	+1	+2/3	+1/3	0	−1/3	−2/3	−1
ANTILEPTONS	$\bar{e}$			$\bar{\nu}_e$			
QUARKS		u			d		
ANTIQUARKS			$\bar{d}$			$\bar{u}$	
LEPTONS				ν_e			e

FIRST GENERATION of quarks and leptons forms an orderly pattern when the particles are arranged according to their electric charge. All values of charge from +1 to −1 in intervals of 1/3 are represented. All colored particles have fractional charge and all colorless ones have integral charge. The pattern is an arbitrary feature of the standard model, where charge and color are independent, but it might have some explanation if quarks and leptons are composite.

among them. The cause of parsimony would be served, for example, if two of the forces could be consolidated, as electricity and magnetism were, or if one force could be made a residue of another, as the strong force was made a residue of the color force. Ironically, it may turn out that a simplification of this kind can be attained only by introducing still more forces.

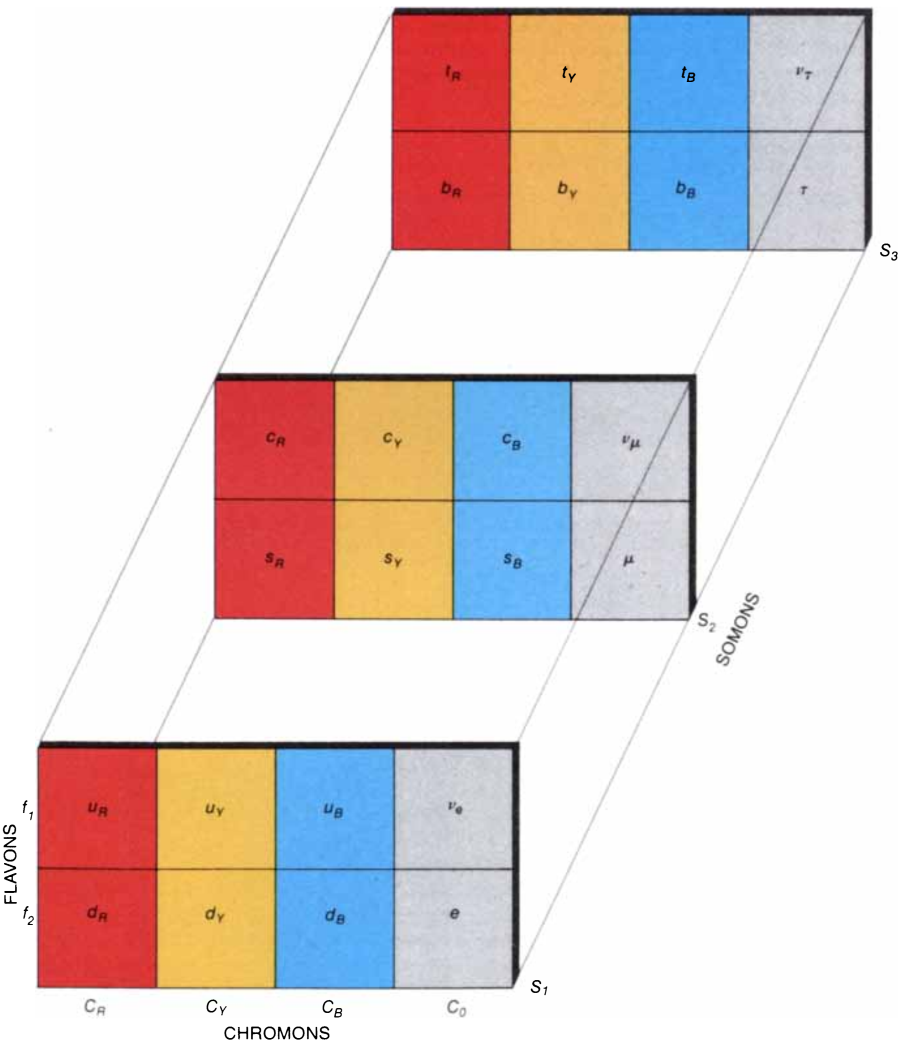
A theory that "goes beyond" the standard model need not contradict or invalidate it. The standard model may emerge as a very good approximation of the deeper theory. The standard model gives a remarkably successful description of all phenomena at distances no smaller than about 10^{-16} centimeter. A deeper theory should therefore focus on events at a still smaller scale. If there are new constituents to be discovered, they must exist within such minuscule regions of space. If there are new forces, their action must be effective only at a distance of less than 10^{-16} centimeter, either because the force is inherently short-range (following the example of the weak force) or because it is subject to some form of confinement (as the color force is).

The search for a theory beyond the standard model was launched almost 10 years ago, and by now several directions have been explored. One promising direction has led to the models known as grand unified theories, which incorporate the electromagnetic, color and weak forces into one fundamental force. The essential idea is to put all the quarks and leptons that make up one generation into a single family; new gauge bosons are then postulated to mediate interactions between the colored quarks and the colorless leptons. The theories account for the regularities noted in the distribution of electric charge and explain the exact commensurability of the quark and lepton charges. On the other hand, they do nothing to reduce the number of fundamental constants, they shed no light on the triplication of the generations and they create certain new theoretical difficulties of their own.

There have been several variations on the theme of grand unification. For example, the concept of horizontal symmetry tackles the triplication problem by establishing a symmetry relation among the generations. The mathematically beautiful idea called supersymmetry relates particles whose spin angular momentum is a half-integer (such as the quarks and leptons) to those with integer spin (such as the gauge bosons). The technicolor theory suggests that the Higgs particle of the standard model is a composite object made up of new fundamental entities; they would be bound together by a new force analogous to the color force and called technicolor. Each of these ideas answers some of the ques-

	PREON	ELECTRIC CHARGE	COLOR	GENERATION NUMBER
FLAVONS	f_1	+1/2	COLORLESS	0
	f_2	−1/2	COLORLESS	0
CHROMONS	C_R	+1/6	RED	0
	C_Y	+1/6	YELLOW	0
	C_B	+1/6	BLUE	0
	C_0	−1/2	COLORLESS	0
SOMONS	S_1	0	COLORLESS	1
	S_2	0	COLORLESS	2
	S_3	0	COLORLESS	3

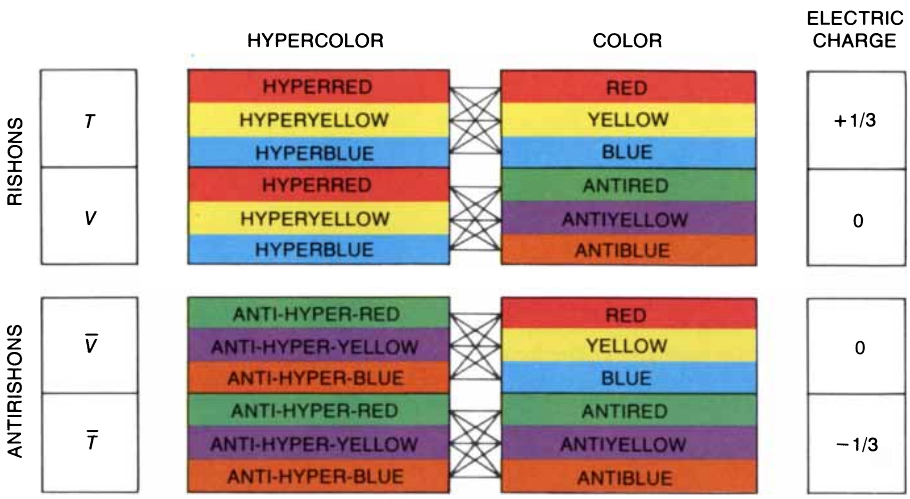
PREON MODEL assigns three properties of quarks and leptons to three groups of hypothetical constituents called flavons, chromons and somons. A quark or a lepton is formed by choosing one preon from each group. The flavons have the primary responsibility for determining electric charge, the chromons determine color and the somons determine generation number. Ideally each kind of preon would carry just one property, but some adjustment is needed to differentiate the fractional electric charges of the quarks from the integral charges of the leptons. In the version of the model shown here the chromons carry electric charge as well as color.



COMBINATIONS OF PREONS give rise to the 24 quarks and leptons of the three generations. For example, the red *s* quark is made up of somon S_2 (signifying that the composite is a second-generation particle), in combination with the red chromon and the negative flavon.

tions that remain open in the standard model. Each idea also fails to answer other questions, raises new difficulties and worsens existing ones, for example by further increasing the number of unrelated arbitrary constants.

In all the above schemes for grand unification it is explicitly assumed that the quarks, the leptons, the photon, the gluons and the weak bosons are the truly fundamental particles of the ultimate theory of nature. The alternative of suggesting that the quarks and leptons are themselves composite is in one sense the most conservative and the least original hypothesis. It is a strategy that has worked before, repeatedly, in going from the atom to the nucleus to the proton to the quark. In another sense the idea of quark and lepton substructure is a most radical proposal. The electron has now been studied for almost a century, and its pointlike nature has been established very well indeed. In the case of the neutrino, which may turn out to be entirely massless, it is even more difficult to imagine an internal structure. The assertion that these particles and the others like them are composites will clearly have to overcome formidable obstacles if it is to have any future.



RISHON MODEL constructs all the quarks and leptons out of just two species of fundamental particles and their antiparticles. The rishons carry both hypercolor, a property associated with the force that binds them to one another, and ordinary color, which they convey to the composite systems they form. One rishon is electrically charged and the other is neutral.

RISHON COMBINATION	PARTICLE	COLOR	ELECTRIC CHARGE
TTT	$\bar{e}$	COLORLESS	+1
TTV	*u*	RED / YELLOW / BLUE	+2/3
TVV	$\bar{d}$	ANTIRED / ANTIYELLOW / ANTIBLUE	+1/3
VVV	ν_e	COLORLESS	0
$\bar{V}\bar{V}\bar{V}$	$\bar{\nu}_e$	COLORLESS	0
$\bar{V}\bar{V}\bar{T}$	*d*	RED / YELLOW / BLUE	−1/3
$\bar{V}\bar{T}\bar{T}$	$\bar{u}$	ANTIRED / ANTIYELLOW / ANTIBLUE	−2/3
$\bar{T}\bar{T}\bar{T}$	*e*	COLORLESS	−1

COMBINATIONS OF RISHONS taken three at a time give a correct accounting of all the quarks and leptons (and antiquarks and antileptons) in any one generation. The pattern of electric charges noted in the standard model, and the apparent relation between fractional charge and color, emerge as natural consequences of the way the rishons combine. All the allowed combinations of three rishons or of three antirishons are neutral with respect to hypercolor.

Offsetting the difficulties of the undertaking are its potential rewards. A fully successful composite model might resolve all the questions left unsettled in the standard model. Such a hypothetical theory would begin by introducing a new set of elementary particles, which I shall refer to generically as prequarks. Ideally there would not be too many of them. Each quark and lepton in the standard model would be accounted for as a combination of prequarks, just as each hadron can be explained as a combination of quarks. The mass of a quark or a lepton would no longer be an arbitrary constant of nature; instead it would be determined by the masses of the constituent prequarks and by the strength of the force that binds the prequarks together. The exact ratios that relate the charge of a quark to that of a lepton would be explained in a similar way: both kinds of composite particles would derive their charges from those of the same constituent prequarks. The entire pattern of quarks and leptons within a generation would presumably reflect some simple rules for combining the prequarks.

The existence of multiple generations might also be explained in a natural way. The quarks and leptons in the higher generations might have an internal constitution similar to that of the corresponding particles of the first generation; the differences could be in the energy and the state of motion of the constituents. Thus the *s* and *b* quarks would be excited states of the *d* quark, and the muon and the tau lepton would be excited states of the electron. Similar excited states are known in all other composite systems, including atoms, nuclei and hadrons. For example, at least a dozen hadrons have been identified in experiments as excited states of the proton; they and the proton itself are all thought to have essentially the same quark composition, namely *uud*.

This imaginary, ideal prequark theory accomplishes everything one might ask of it except for unifying the fundamental forces. Even there some progress is conceivable, since a new force would very likely be introduced to bind the prequarks together; the new force might lead to a new understanding of how the known forces are related. Imagining what a successful model might be like, however, is not at all the same thing as actually constructing a realistic and internally consistent one. So far no one has done it.

What has been lacking is a satisfactory theory of prequark dynamics, a theory that would describe how the prequarks move inside a quark or a lepton and that would enable one to calculate the mass and total energy of the system. As I shall set forth below, there are fundamental obstacles to the formulation of such a theory, although I would submit that they are not insurmountable. In the meantime, lacking any persuasive account of prequark motions, theorists have nonetheless been exploring the combinatorial possibilities of the prequark idea, that is, they have been examining the ways quarks and leptons might be built up as specific combinations of finer constituents.

In the past few years several dozen composite models have been proposed; they can be classified in perhaps four or five main groups. No single model solves all problems, answers all questions and is widely accepted. It would be unfair to describe only one scheme, but it is impractical to enumerate them all. I shall present a few of the central ideas.

The first explicit model of quark and lepton substructure was proposed in 1974 by Jogesh C. Pati of the University of Maryland at College Park and Salam, who have since returned to the topic several times in collaboration with John Strathdee of the International Centre for Theoretical Physics. It was they who introduced the term prequark, which I have adopted here as a generic name for hypothetical subconstituents of all kinds. The specific elementary particles of the model devised by Pati and Salam I shall call preons, which is another term of their invention.

The rationale for the preon model begins with the observation that every quark and lepton can be identified unambiguously by listing just three of its properties: electric charge, color and generation number. These properties, then, suggest a straightforward way of organizing a set of constituent particles. Three families of preons are needed. In one family the preons carry electric charge, in another they carry color and in the third they have some property that determines generation number. A given quark or lepton is assembled by selecting exactly one preon from each family.

The preons that determine generation number are called somons, from the Greek *soma,* meaning body, because they have a dominant influence on the mass of the composite system. Since there are three generations of quarks and leptons, there must be three somons. The color of the composite system is determined by preons called chromons; there are four of them, one with the color red, one yellow, one blue and one colorless. The remaining family of preons, which is assigned the role of defining electric charge, needs to have only two members in order for every quark and lepton to be uniquely identified. These last preons have been given the name flavons, after flavor, the whimsical term for whatever property it is that distinguishes the u quark from the d quark, the c from the s, the neutrino from the electron and so on.

In the preon model the classification of a composite particle follows directly from its complement of preons. All leptons, for example, are distinguished by a colorless chromon, and all first-generation particles must obviously have a first-generation somon. In the allocation of electric charge, however, a complication arises. If there are only two flavons and if they are the sole carriers of electric charge, not all the charge values observed in nature can be reproduced. The u quark and the neutrino, for example, must have the same charge (because they include the same flavon), and so must the d quark and the electron. The problem can be solved in any of several ways. In one scheme electric charge is assigned to both the flavons and the chromons, and the total charge of a composite particle is equal to the sum of the two values. Models of this kind can be made to yield the correct charge states, but only by abandoning the principle of having each kind of preon carry just one property.

Another troublesome feature of the preon model is the requirement that composites be formed only by drawing one preon from each family. Why are there no particles made up of three chromons, say, or of two somons and a flavon? The exotic properties of such particles would make them quite conspicuous. It seems likely that if they existed, they would have been detected by now.

Many variations of the preon model have been proposed by other physicists, using the same basic idea but slightly different sets of preons. Notable among the variations are the models suggested by Hidezumi Terazawa, Yoichi Chikashige and Keiichi Akama of the University of Tokyo and by O. Wallace Greenberg and Joseph Sucher of the University of Maryland.

Perhaps the simplest model of quark and lepton structure is the rishon model, which I proposed in 1979. A similar idea was put forward at about the same time by Michael A. Shupe of the University of Illinois at Urbana-Champaign. The model has since been further developed and studied in great detail by Nathan Seiberg and me at the Weizmann Institute of Science in Rehovot. The model postulates just two species of fundamental building blocks, called rishons. (*Rishon* is the Hebrew adjective meaning first or primary.) One rishon has an electric charge of $+1/3$ and the other is electrically neutral. I designate them respectively T and V, for *Tohu Vavohu,* Hebrew for "formless and void," the description of the initial state of the universe given in the first chapter of Genesis. The complementary antirishons have charges of $-1/3$ and zero and are designated $\overline{T}$ and $\overline{V}$.

The model has one simple rule for constructing a quark or a lepton: any three rishons can be assembled to form a composite system, or any three antirishons, but rishons and antirishons cannot be mixed in a single particle. The rule gives rise to 16 combinations, which reproduce exactly the properties of the 16 quarks, antiquarks, leptons and antileptons in the first generation. In other words, every quark and lepton in the first generation corresponds to some allowed combination of rishons or antirishons. (In this system of classification each color is counted separately.)

The pattern of quark and lepton charges is generated as follows. The TTT combination, with rishon charges of $1/3 + 1/3 + 1/3$, has a total charge of $+1$ and therefore corresponds to the positron; similarly, $\overline{TTT}$ has a total charge of -1 and is identified with the electron. The VVV and $\overline{VVV}$ combinations are both electrically neutral and represent the neutrino and the antineutrino respectively. The remaining allowed combinations yield fractionally charged quarks. TTV, with a charge of $+2/3$, is the u quark, and TVV, with a charge of $+1/3$, is the $\bar{d}$ antiquark. The analogous antirishon states $\overline{VVT}$ and $\overline{VTT}$ correspond to the d quark and the $\bar{u}$ antiquark.

The model also accounts successfully for the color of the composite systems. A T rishon can have any of the three colors red, yellow and blue, whereas a V rishon has an anticolor. Combinations such as $\overline{TTT}$ and VVV, which designate leptons, can be made colorless since they can include one rishon in each color or one in each anticolor. The other combinations, which yield quarks, must have a net color. For example, a TTV state might have the rishon colors red, blue and antiblue; the antiblue would cancel the blue, leaving the system with a net color of red. In this way the connection between color and electric charge, which was apparent but unexplained in the standard model, is readily understood. Because of the way electric charge and color are allotted to the rishons, all composite systems with fractional charge turn out to be colored, and all systems with an integer charge can be made colorless.

Other regularities of the standard

model also lose their air of mystery when rishons are introduced. Consider the hydrogen atom, made up of a proton and an electron, or in terms of quarks and leptons two *u* quarks, a *d* quark and an electron. The total rishon content of the quarks is four *T*'s, one $\overline{T}$, two *V*'s and two $\overline{V}$'s. The electric charge of the $\overline{T}$ cancels the charge of one *T* rishon, and the *V*'s and $\overline{V}$'s also cancel (they have no charge in any case), leaving the proton with a net charge equal to that of a *TTT* system. The electron's rishon content is just the opposite: $\overline{T}\overline{T}\overline{T}$. Thus it is evident why the proton and the electron have charges of equal magnitude and why the hydrogen atom is neutral: the ultimate sources of the charge are pairs of matched particles and antiparticles.

The rishon model and many other models that explain the pattern of the first generation have difficulty accounting for the second and third generations. It would seem that such models lend themselves well to the scheme of forming each particle in the higher generations as an excited state of the corresponding particle in the first generation. The simplest idea would be to describe the muon, for example, as having the same prequark constituents as the electron, but in the muon the prequarks would have some higher-energy configuration. It is an elegant idea but, regrettably, it appears to be unworkable. The scheme implies differences in energy between the successive excited states that are much larger than the actual differences. The flaw is a fundamental one, and there seems to be no remedy.

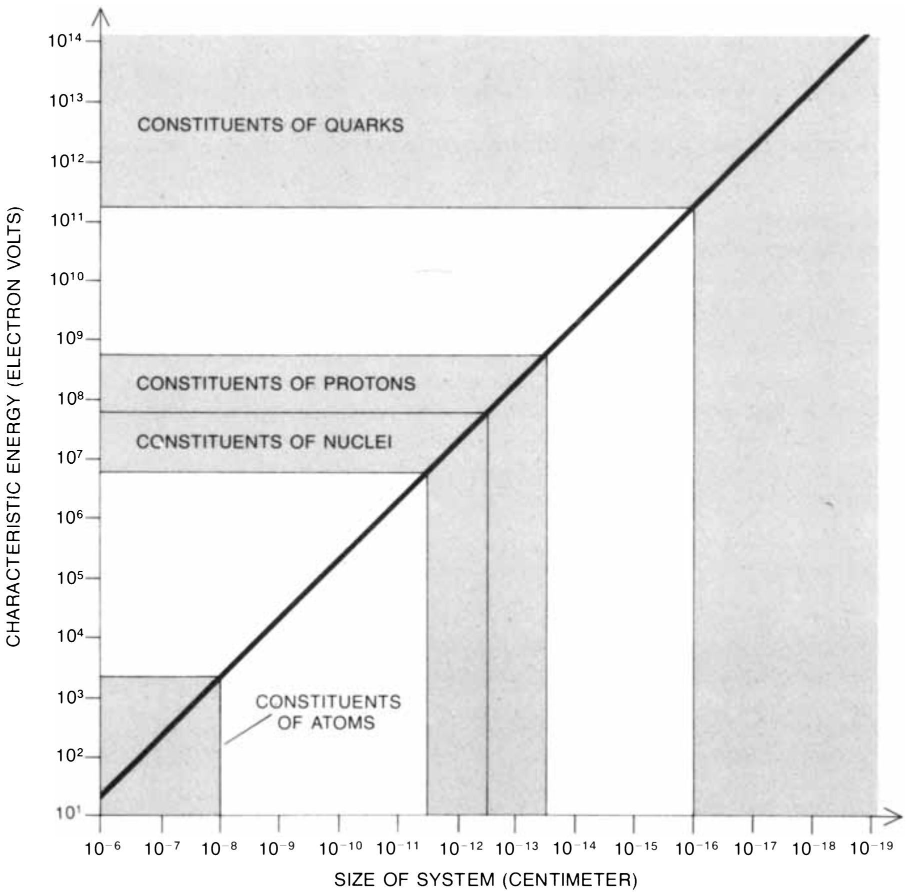
Other possible mechanisms for creating multiple generations have been considered. Several physicists have suggested that the higher-generation relatives of a given state might be created by adding a Higgs particle, the "extra" particle associated with the weak bosons in the standard model. Because a Higgs particle has no electric charge or color or even spin angular momentum, adding one to a composite system would alter only the mass. Hence an electron might be converted into a muon by adding one Higgs particle or into a tau by adding two or more Higgs particles. Seiberg and I have proposed another possible mechanism: a higher-generation particle could be formed by the addition of pairs of prequarks and antiprequarks. All charges and other properties must cancel in such a pair, and so again only the mass would be affected.

These ideas are currently at the stage of unrestrained speculation. No one knows what distinguishes the three generations from one another, or why there are three or whether there may be more. No explanation can be given of the mass difference between the generations. In short, the triplication of the generations is still a major unsolved puzzle.

A third kind of substructure model deserves mention. It tries to relate the possibility of quark and lepton structure to another fundamental problem: understanding the relativistic quantum theory of gravitation. Ideas of this kind have been explored by John Ellis, Mary K. Gaillard, Luciano Maiani and Bruno Zumino of CERN. One approach to their ideas is to consider the distances at which prequarks interact: the experimental limit is less than 10^{-16} centimeter, but the actual distance could be several orders of magnitude smaller still. At about 10^{-34} centimeter the gravitational force becomes strong enough to have a significant effect on individual particles. If the scale of the prequark interactions is this small, gravitation cannot be neglected. Ellis, Gaillard, Maiani and Zumino have outlined an ambitious program that aims to unify all the forces, including gravitation, in a scheme that treats not only the quarks and leptons but also the gauge bosons as composite particles. Like other composite models, however, this one has serious flaws.

Any prequark model, regardless of its details, must supply some mechanism for binding the prequarks together. There must be a powerful attractive force between them. One strategy is to postulate a new fundamental force of nature analogous in its workings to the color force of the standard model. To emphasize the analogy the new force is called the hypercolor force and the carrier fields are called hypergluons. The prequarks are assumed to have hypercolor, but they combine to form hypercolorless composite systems, just as quarks have ordinary color but combine to form colorless protons and neutrons. The hypercolor force presumably also gives rise to the property of confinement, again in analogy to the color force. Hence all hypercolored prequarks would be trapped inside composite particles, which would explain why free prequarks are not seen in experiments. An idea of this kind was first proposed by 't Hooft, who studied some of its mathematical implications but also expressed doubt that nature actually follows such a path.

The typical radius of hypercolor con-



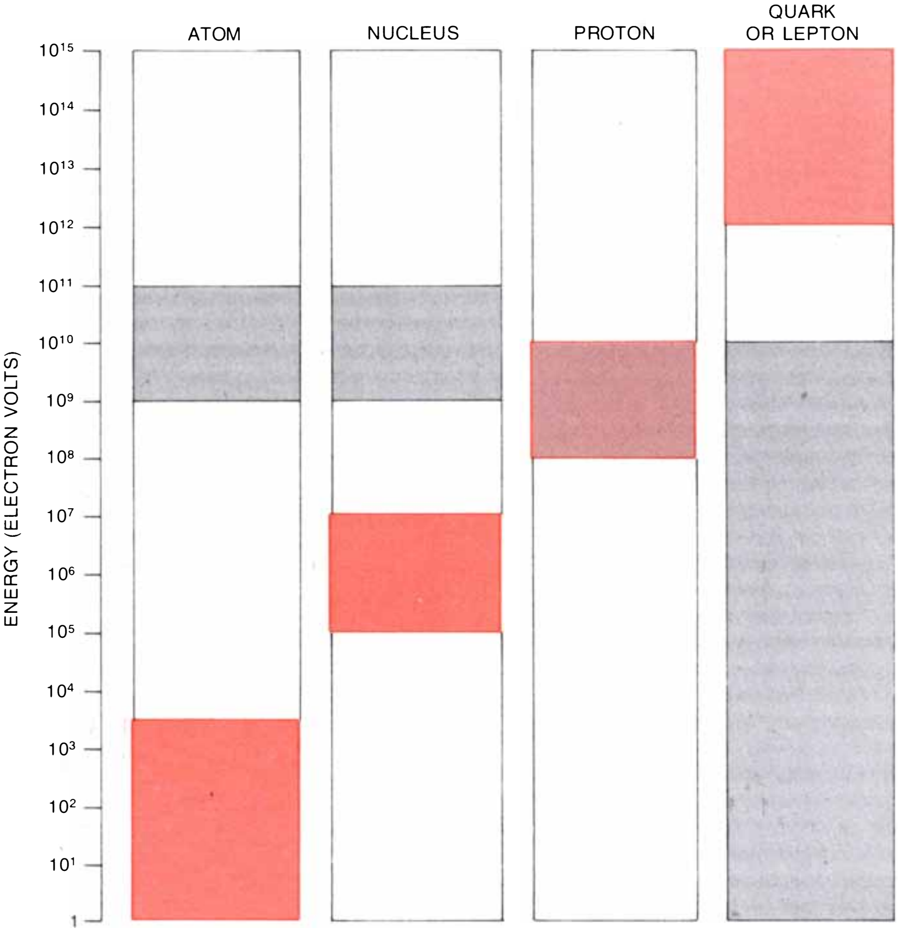
SIZE AND ENERGY have a reciprocal relation in the quantum theory, indicating that the constituents of composite quarks and leptons must have an exceedingly high kinetic energy. The size of an atom implies that its constituents can have energies ranging from a few electron volts to a few thousand. (An electron volt is the energy gained by an electron accelerated through a potential difference of one volt.) In a nucleus the protons and neutrons move with an energy of several million electron volts, and in a proton or a neutron the quarks have energies of several hundred million electron volts. Any constituents of quarks and leptons must be confined to a radius of less than 10^{-16} centimeter, and possibly much less. As a result the kinetic energy of the hypothetical prequarks can be no less than a few hundred billion electron volts.

finement must be less than 10^{-16} centimeter. Only when matter is probed at distances smaller than this would it be possible to see the hypothetical prequarks and their hypercolors. At a range of 10^{-14} or 10^{-15} centimeter hypercolor effectively disappears; the only objects visible at this scale of resolution (the quarks and leptons) are neutral with respect to hypercolor. At a range of 10^{-13} centimeter ordinary color likewise fades away, and the world seems to be made up entirely of objects that lack both color and hypercolor: protons, neutrons, electrons and so on.

The notion of hypercolor is well suited to a variety of prequark models, including the rishon model. In addition to their electric charge and color the rishons are assumed to have hypercolor and the antirishons to have antihypercolor. Only combinations of three rishons or three antirishons are allowed because only those combinations are neutral with respect to hypercolor. A mixed three-particle system, such as $TT\bar{T}$, cannot exist because it would not be hypercolorless. The assignment of hypercolors thereby explains the rule for forming composite rishon systems. Similar rules apply in other hypercolor-based prequark models.

If the aim of a prequark model is to simplify the understanding of nature, postulating a new basic force does not seem very helpful. In the case of hypercolor, however, there may be some compensation. Consider the neutrino: it has neither electric charge nor color, only weak charge. According to the standard model, two neutrinos can act on each other only through the short-range weak force. If neutrinos are composites of hypercolored prequarks, however, there could be an additional source of interactions between neutrinos. When two neutrinos are far apart, there are practically no hypercolor forces between them, but when they are at close range, the hypercolored prequarks inside one neutrino are able to "see" the inner hypercolors of the other one. Complicated short-range attractions and repulsions are the result. The mechanism, of course, is exactly the same as the one that explains the molecular force as a residue of the electromagnetic force and the strong force as a residue of the color force.

The conclusion may also be the same. Seiberg and I, and independently Greenberg and Sucher, were the first to suggest that the short-range weak force may actually be a residual effect of the hypercolor force. According to this hypothesis, the weak bosons W^+, W^- and Z^0 must also be composite objects, presumably made up of certain combinations of the same prequarks that compose the quarks and leptons. If this idea is confirmed, the list of fundamental forces will still have four entries: gravitation, electromagnetism, color and hypercolor. It should be noted, however, that all these forces are long-range ones; the short-range molecular, strong and weak forces will have lost their fundamental status.



MISMATCH OF ENERGY AND MASS makes it difficult to devise a theory of how prequarks might move and interact. In an atom or a nucleus the kinetic energy of the constituents (*color*) is much less than the total mass of the system (*gray*). In a proton the two quantities are of comparable magnitude. In a composite quark, however, the energy of the prequarks greatly exceeds the total mass. Indeed, compared with the kinetic energy, the mass is essentially zero. Somehow virtually all mass is canceled, a development that is unlikely to be accidental.

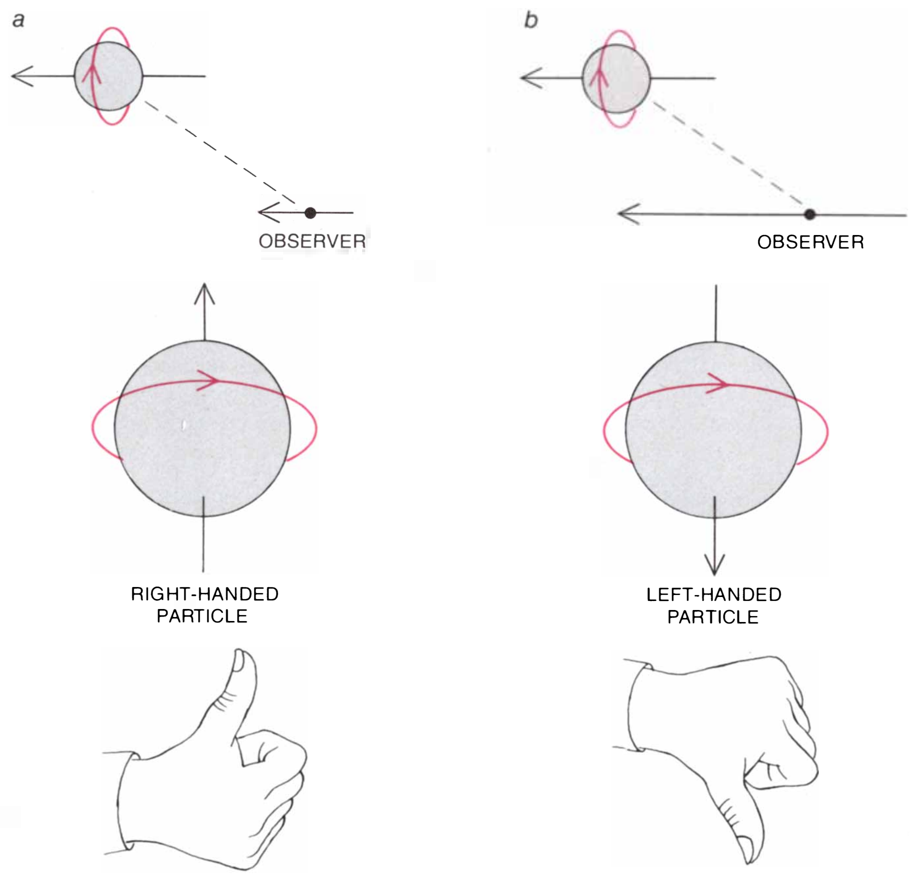
For now hypercolor remains a conjecture, and so does the notion of explaining the weak force as a residue of the hypercolor force. It may yet turn out that the weak force is fundamental. A careful measurement of the mass, lifetime and other properties of the weak bosons should provide important clues in this matter.

Hypercolor is not the only candidate for a prequark binding force. Another interesting possibility was suggested by Pati, Salam and Strathdee. Instead of introducing a new hypercolor force, they borrowed an idea that has long been familiar, namely the magnetic force, and adapted it to a new purpose. An ordinary magnet invariably has two poles, which can be thought of as opposite magnetic charges. For 50 years there have been theoretical reasons for supposing there could also be isolated magnetic charges, or monopoles. Pati, Salam and Strathdee have argued that the prequarks could be particles with charges resembling both magnetic and electric charges. If they are, the forces binding them may be of a new and interesting origin.

None of the ideas I have just described constitutes a theory of prequark dynamics. Indeed, there is a serious impediment to the formulation of such a theory; it is the requirement that the prequarks be exceedingly small. The most stringent limit on their size is set indirectly by measurements of the magnetic moment of the electron, which agree with the calculations of quantum electrodynamics to an accuracy of 10 significant digits. In the calculations it is assumed that the electron is pointlike; if it had any spatial extension or internal structure, the measured value would differ from the calculated one. Evidently any such discrepancy can at most affect the 11th digit of the result. It is this constraint that implies the characteristic distance scale of the electron's internal structure must be less than 10^{-16} centimeter. Roughly speaking, that is the maximum radius of an electron, and any prequarks must stay within it. If they strayed any wider, their presence would already have been detected.

Why should the small size of the electron inhibit speculation about its internal structure? The uncertainty principle establishes a reciprocal relation between the size of a composite system and the kinetic energy of any components moving inside it. The smaller the system, the larger the kinetic energy of the constituents. It follows that the prequarks must have enormous energy: more than 100 GeV (100 billion electron volts), and possibly much more. (One electron volt is the energy acquired by an electron when it is accelerated through a potential difference of one volt.) Because mass is fundamentally equivalent to energy, it can be measured in the same system of units. The mass of the electron, for example, is equivalent to .0005 GeV. There is a paradox here, which I call the energy mismatch: the mass of the composite system (if it is indeed composite) is much smaller than the energy of its constituents.

The oddity of the situation can be illuminated by considering the relations of mass and kinetic energy in other composite systems. In an atom the kinetic energy of a typical electron is smaller than the mass of the atom by many orders of magnitude. In hydrogen, for example, the ratio is roughly one part in 100 million. The energy needed to change the orbit of the electron and thereby put the atom into an excited state is likewise a negligible fraction of the atomic mass. In a nucleus the kinetic energy of the protons and neutrons is also small compared with the nuclear mass, but it is not completely negligible. The motion of the particles gives them an energy equivalent to about 1 percent of the system's mass. The energy needed to create an excited state is also about 1 percent of the mass.



CHIRAL SYMMETRY offers a possible explanation of the "miraculous" cancellation of mass in quarks and leptons. Chirality, or handedness, describes the relation of a particle's spin angular momentum to its direction of motion. Suppose an observer is overtaken by a faster-moving electron (*a*). From the observer's point of view the electron obeys a right-hand rule: When the fingers of the right hand curl in the same direction as the spin, the thumb gives the direction of motion. If the observer speeds up, however, so that he overtakes the electron (*b*), the handedness of the particle changes. In the observer's frame of reference the electron is now approaching instead of receding, but its spin direction has not changed; as a result its motion is described by a left-hand rule. Chirality, therefore, is not conserved. There is one kind of particle to which this argument cannot be applied, namely a massless particle, which must always move with the speed of light. No observer can move faster than a massless particle, and so its handedness is an invariant property. If a theory of prequarks had a chiral symmetry, in which handedness must be conserved, the low mass of the quarks and leptons might not be accidental. They would have to be virtually massless for the chiral symmetry to be maintained.

With the proton and its quark constituents the energy-mass relation begins to get curious. From the effective radius of the proton the typical energy of its component quarks can be calculated; it turns out to be comparable to the mass of the proton itself, which is a little less than 1 GeV. The energy that must be invested to create an excited state of the quark system is of the same order of magnitude: the hadrons identified as excited states of the proton exceed it in mass by from 30 to 100 percent. Nevertheless, the ratio of kinetic energy to total mass is still in the range that seems intuitively reasonable. Suppose one knew only the radius of the proton, and hence the typical energy of whatever happens to be inside it, and one were asked to guess the proton's mass. Since the energy of the constituents is generally a few hundred million electron volts, one would surely guess that the total mass of the system is at least of the same order of magnitude and possibly greater. The guess would be correct.

For the atom, the nucleus and the proton, then, the mass of the system is at least as large as the kinetic energy of the constituents and in some cases is much larger. If quarks and leptons are composite, however, the relation of energy to mass must be quite different. Since the prequarks have energies well above 100 GeV, one would guess that they would form composites with masses of hundreds of GeV or more. Actually the known quarks and leptons have masses that are much smaller; in the case of the electron and the neutrinos the mass is smaller by at least six orders of magnitude. The whole is much less than the sum of its parts.

The high energy of the prequarks is also what spoils the idea of viewing the higher generations of quarks and leptons as excited states of the same set of prequarks that form the first-generation particles. As in the other composite systems, the energy needed to change the orbits of the prequarks should be of the same order of magnitude as the kinetic energy of the constituents. One would therefore expect the successive generations to differ in mass by hundreds of GeV, whereas the actual mass differences are as small as .1 GeV.

At this point one might well adopt the view that the energy mismatch cannot be accepted, indeed that it simply demonstrates the elementary and structureless nature of the quarks and leptons. Many physicists hold this view. The energy mismatch, however, contradicts no basic law of physics, and I would argue that the circumstantial evidence for quark and lepton compositeness is sufficiently persuasive to warrant further investigation.

What is peculiar about the quark and lepton masses is not merely that they are small but that they are virtually zero when measured on the energy scale defined by their constituents' energy. In other composite systems a small amount of mass is "lost" by being converted into the binding energy of the system. The total mass of a hydrogen atom, for example, is slightly less than that of an isolated proton and electron; the difference is equal to the binding energy. In a nucleus this "mass defect" can reach a few percent of the total mass. In a quark or a lepton, it seems, the entire mass of the system is canceled almost exactly. Such a "miraculous" cancellation is certainly not impossible, but it seems most unlikely to happen by accident. Similar large cancellations are known elsewhere in physics, and they have always been found to result from some symmetry principle or conservation law. If there is to be any hope of constructing a theory of prequark dynamics, it is essential to find such a symmetry in this case.

SPONTANEOUS SYMMETRY BREAKING is a mechanism that could spoil a prequark theory even if it has a chiral symmetry. Both of the physical systems shown here—a simple trough and a trough with a bump in the bottom—can be described as symmetrical in the sense that exchanging left and right leaves the system unaltered. For the simple trough the system remains symmetrical when a ball is put in the trough; the ball comes to rest in the center, so that exchanging left and right still has no effect. In the trough with a bump, however, the ball takes up a position on one side or the other, and the symmetry is inevitably broken. Similarly, a prequark theory that has a chiral symmetry might nonetheless give rise to composite systems that do not observe the symmetry. Showing that a chiral symmetry can definitely remain unbroken is currently the principal challenge in formulating a theory of how prequarks move.

There is a likely candidate: chiral symmetry, or chirality. The name is derived from the Greek word for hand, and the symmetry has to do with handedness, the property defined by a particle's spin and direction of motion. Like other symmetries of nature, chiral symmetry has a conservation law associated with it, which gives the clearest account of what the symmetry means. The law states that the total number of right-handed particles and the total number of left-handed ones can never change.

In the ordinary world of protons, electrons and similar particles handedness or chirality clearly is not conserved. A violation of the conservation law can be demonstrated by a simple thought experiment. Imagine that an observer is moving in a straight line when he is overtaken by an electron. As the electron recedes from him he notes that its spin and direction of motion are related by a right-hand rule. Now suppose the observer speeds up, so that he is overtaking the electron. In the observer's frame of reference the electron seems to be approaching; in other words, it has reversed direction. Because its spin has not changed, however, it has become a left-handed particle.

There is one kind of particle to which this thought experiment cannot be applied: a massless particle. Because a massless particle must always move with the speed of light, no observer can ever go faster. As a result the handedness of a massless particle is an invariant property, independent of the observer's frame of reference. Furthermore, it can be shown that none of the known forces of nature (those mediated by the photon, the gluons and the weak bosons) can alter the handedness of a particle. Thus if the world were made up exclusively of massless particles, the world could be said to have chiral symmetry.

Chiral symmetry is the root of an idea that might conceivably account for the small mass of the quarks and leptons. The argument runs as follows. If the prequarks are massless particles, if they have a spin of 1/2 and if they interact with one another only through the exchange of gauge bosons, any theory describing their motion is guaranteed to have a chiral symmetry. If the massless prequarks then bind together to form composite spin-1/2 objects (namely the quarks and leptons), the chiral symmetry might ensure that the composite particles also remain massless compared with the huge energy of the prequarks inside them. Hence the small mass of the quarks and leptons is not an accident. They must be essentially massless with respect to the energy of their constituents if the chiral symmetry of the theory is to be maintained.

The crucial step in this argument is the one extending the chiral symmetry from a world of massless prequarks to one made up of composite quarks and leptons. It is essential that the symmetry of the original physical system survive in and be respected by the composite states formed out of the massless constituents. It may seem self-evident that if a theory is symmetrical in some sense, the physical systems described by the theory must exhibit that symmetry; actually, however, the spontaneous breaking of symmetries is commonplace. A familiar example is the roulette wheel. A physical theory of the roulette wheel would show it is completely symmetrical in the sense that each slot is equivalent to any other slot. The physical system formed by putting a ball in the roulette wheel, however, is decidedly asymmetrical: the ball invariably comes to rest in just one slot.

In the standard model it is the spontaneous breaking of a symmetry that makes the three weak bosons massive and leaves the photon massless. The theory that describes these gauge bosons is symmetrical, and in it the four bosons are essentially indistinguishable, but because of the symmetry breaking the physical states actually observed are quite different. Chiral symmetries are notoriously susceptible to symmetry breaking. Whether the chiral symmetry of prequarks breaks or not when the prequarks form composite objects can be determined only with a detailed understanding of the forces acting on the prequarks. For now that understanding does not exist. In certain models it can be shown that a chiral symmetry does exist but is definitely broken. No one has yet succeeded in constructing a composite model of quarks and leptons in which a chiral symmetry is known to remain unbroken. Neither the preon model nor the rishon model succeeds in solving the problem. The task is probably the most difficult one facing those attempting to demonstrate that quarks and leptons are composite.

If a consistent prequark theory can be worked out, it will still have to pass the test of experiment. First, it is important to establish in the laboratory whether or not quarks and leptons have any internal structure at all. If they do, experiments might then begin to discriminate among the various models. The experiments will have to penetrate the unknown realm of distances smaller than 10^{-16} centimeter and energies higher than 100 GeV. There are two basic ways to explore this region: by doing experiments with particles accelerated to very high energy and by making precise measurements of low-energy quantities that depend on the physics of events at very small distances.

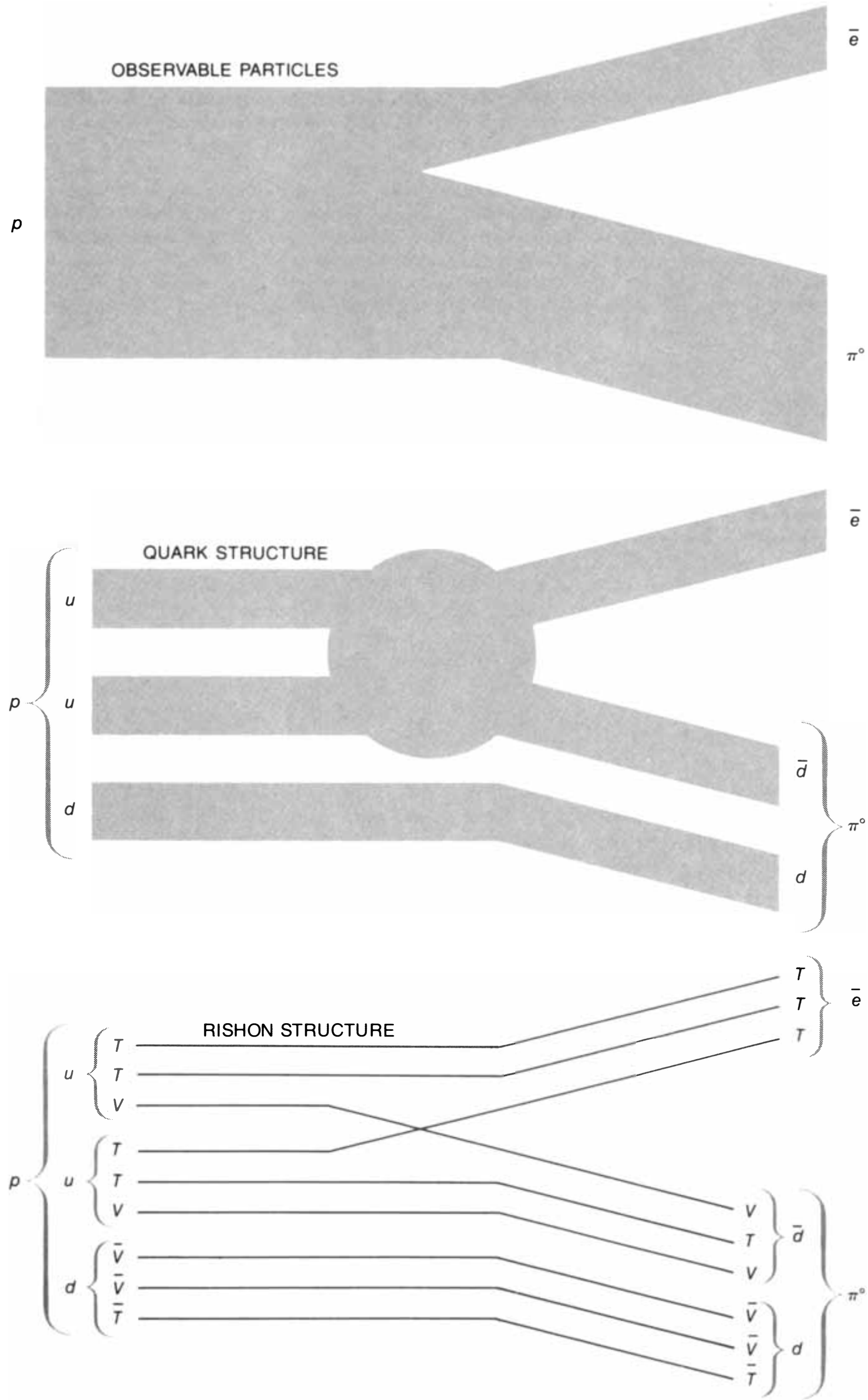
Experiments of the first kind include the investigation of the weak bosons and the search for the Higgs particles of the standard model. When such particles can be made in sufficient numbers, a careful look at their properties should reveal much about the physics of very small distances. New accelerators now being planned or built in the U.S., Europe and Japan are expected to yield detailed information about the weak bosons and will also continue the ongoing investigation of the quarks and leptons themselves.

Equally interesting are the high-precision, low-energy experiments. One of these is the search for the decay of the proton, a particle that is known to have an average lifetime of at least 10^{30} years. Several experiments are now monitoring large quantities of matter, incorporating substantially more than 10^{30} protons, in an attempt to detect the signals emitted when a proton disintegrates. None of the forces of the standard model can induce such an event, but none of the rules of the standard model absolutely forbids it. Both the grand unified theories and the prequark models, on the other hand, include mechanisms that could convert a proton into other particles that would ultimately leave behind only leptons and photons. If the decay is detected, its rate and the pattern of decay products could offer an important glimpse beyond the standard model.

There is similar interest in the hypothetical process in which a muon emits a photon and is thereby converted into an electron. Again none of the forces of the standard model can bring about an event of this kind, but again too no fundamental law forbids it. Some of the composite models allow the transition and others do not, so that a search for the process might offer a means of choosing among the models. Experiments done up to now put a limit of less than one in 10 billion on the probability that any given muon will decay in this way. Detection of such events and a determination of their rate might illuminate the mysterious distinction between the generations.

A third class of precision experiments are those that continue to refine the measurement of the magnetic moment of the electron and of the muon. Further improvements can be expected both in experimental accuracy and in the associated calculations of quantum electrodynamics. If the results continue to agree with the predictions of the standard model, the limit on the possible size of any quark and lepton substructure will become remoter. If a discrepancy between theory and experiment is detected, it will represent a strong hint that quarks and leptons are not elementary.

It may well be a decade or two before the next level in the structure of matter comes clearly into view (if, again, there is another level). What is needed is a sound theoretical picture, one that is self-consistent, that agrees with all experiments and that is simple enough to explain all the features of the standard model in terms of a few principles and a few fundamental particles and forces. The correct picture, whether it is a grand unified theory or a composite model of the quarks and leptons, may already exist in some embryonic form. On the other hand, it is also possible the correct theory will emerge only from some totally new idea. In the words of Niels Bohr, it may be that our present ideas "are not sufficiently crazy to be correct."



DECAY OF THE PROTON is a conjectured event that might be interpreted as experimental evidence for a grand unified theory or for a model of quark substructure. In one form of decay the proton would be observed to disintegrate into a positron ($\bar{e}$) and a neutral pion (π^0). The event can be understood in terms of the proton's quark constituents: an interaction of the two *u* quarks converts one of them into a positron and the other into a $\bar{d}$ antiquark; the latter combines with the remaining *d* quark of the proton to form the neutral pion. Grand unified theories suggest that the interaction of the *u* quarks is mediated by a new force of nature. The rishon model provides an alternative explanation: the two *u* quarks merely exchange a *T* and a *V* rishon.

Reshaping the future of film.

From fat to flat. The goal many of us strive for with laborious push-ups and strenuous diets, scientists have achieved in Kodak laboratories. The "bodies" they have streamlined, however, are those of silver halide grains. More than a beauty treatment, this process of reshaping enabled us to position grains for maximum light-catching ability, and expand their sensitizing characteristics. It's this advance in silver halide emulsion technology which made it possible for us to produce a true 1000-speed film with a balance of speed, grain, and sharpness previously unattainable.

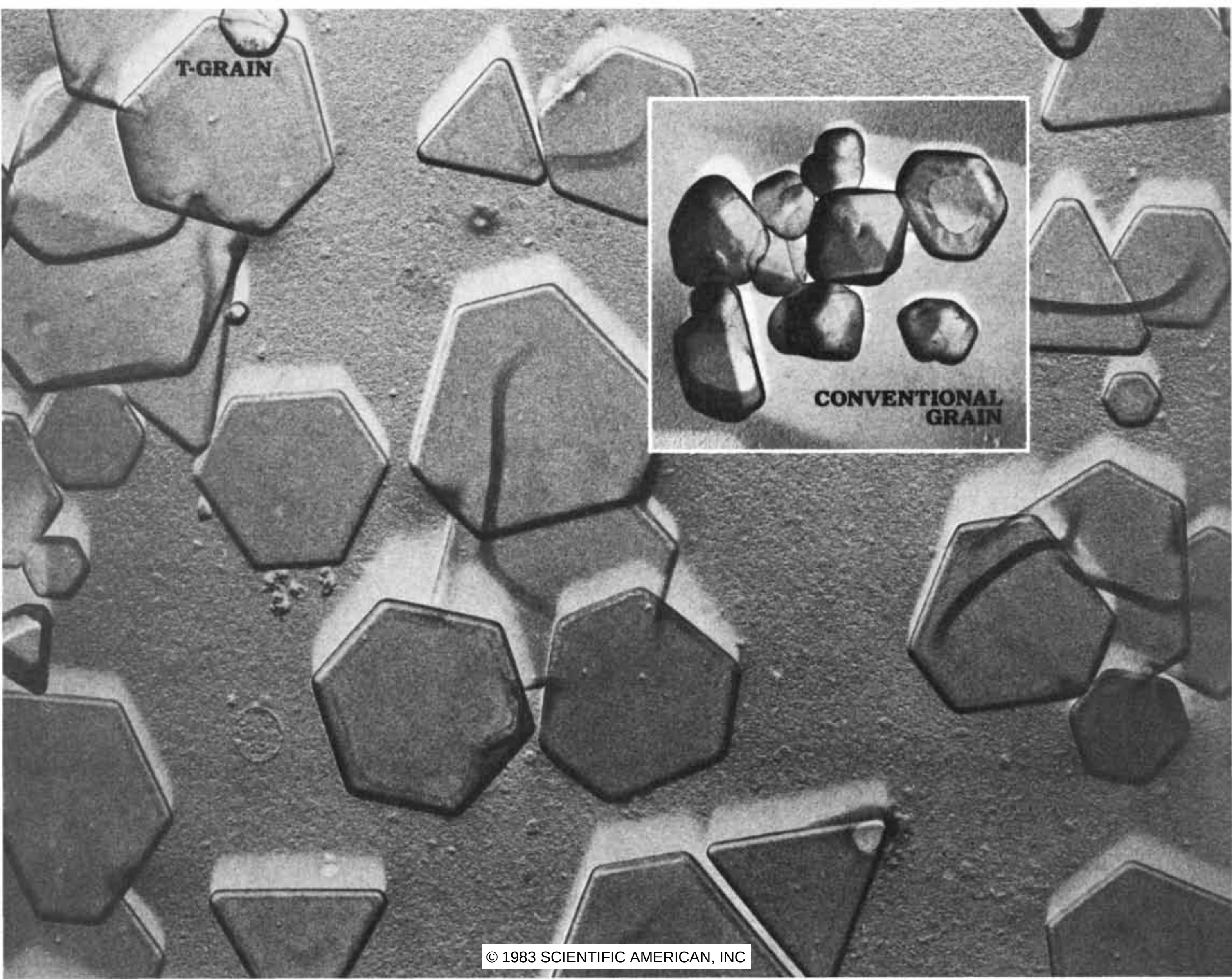
The structure of new Kodacolor VR 1000 film is a modification of the inverted-layer format first used in Kodacolor 400 film. This concept places the fastest light-sensitive layers closer to the camera lens. The magenta and cyan layers are repositioned to form a fast-magenta, fast-cyan, slow-magenta, slow-cyan sequence. In the new emulsion, the yellow-filter layer is eliminated, and the protective overcoat and ultraviolet-filter layers are combined into one, without any loss in color quality.

The resultant faster film can be exposed by daylight, electronic flash, or tungsten, which provides photographers with greater picture-taking capability, particularly when recording fast-action scenes or shooting in low-light conditions.

While reshaping the light-sensitive substance of film is a significant accomplishment in itself, we view this new technology as a stepping-stone. It may well reshape the future of film and enable photography to do hundreds of different jobs better than before.

For more information on Kodak's fastest color print film ever and the new emulsion technology behind it...send your request to: Eastman Kodak Company, Department GBSA-8, 343 State St., Rochester, NY 14650.

Kodak. Where technology anticipates need.



Hot Springs on the Ocean Floor

A sign of the volcanic activity that mints new oceanic crust, they nourish strange forms of life and lay down great deposits of ore. They also help to explain the chemical content of the sea

by John M. Edmond and Karen Von Damm

Hot springs are common at the bottom of the ocean along the submarine ridges where large areas of the surface of the earth are being driven apart and new oceanic crust is rising into place. One might have made that prediction soon after the formulation of the theory of plate tectonics, which explains the significance of the ridges. Indeed, the prediction was made, by J. W. Elder of the University of Manchester. Now, two decades later, it is being confirmed. Cameras, sensors and research submarines operating at ocean depths approaching three kilometers are finding large numbers of hot springs, and the water they vent into the sea is proving to have remarkable consequences.

First, the chemical content of the vent water, which originates in a complex set of reactions between seawater and the hot, newly minted crustal rock, is proving to nourish a chain of living organisms extending from bacteria to clams and giant tube-dwelling worms. Second, the hydrothermal reactions are proving to be the main source of the metal-rich sediments and nodules that carpet the floor of the ocean. The exploitation of these deposits is now under worldwide debate. Other ore deposits, which lie on the continents and have been exploited since ancient times, are proving to have been emplaced there by plate-tectonic motions after hydrothermal activity had manufactured them under the sea. Third, the chemical content of the vent water is proving to be an important determinant of the chemical composition of the oceans themselves.

In retrospect the place of hot springs in the theory of plate tectonics is quite straightforward. In the theory the surface of the earth is viewed as a set of large, rigid plates that are in constant motion with respect to one another. The plates themselves are mostly the crust underlying the oceans, with great rafts of lighter material, the continents, embedded in them. At the margins where plates collide, old crust is being consumed. Such margins are trenches where plates plunge into the plastic interior of the earth's mantle. At the margins where plates diverge, new crust is being created. These latter margins are almost entirely under the oceans. They consist of linear features on the order of 100 kilometers long, offset in a zigzag pattern by faults a few kilometers long.

The Midocean Ridges

Magma (molten rock produced by the partial melting of the mantle at depths of no more than a few hundred kilometers) rises into the linear features. Its temperature as it arrives is some 1,200 degrees Celsius. It cools and solidifies, forming new oceanic crust; then, as the cooling continues, the crust is slowly rifted apart, making way for newer intrusions. The process finds expression in the height of the ocean floor. Hot material expands; hence the intrusion zones are high. They are the midocean ridges, which top off at depths between 2.5 and three kilometers in the sea. Cooling material contracts; hence the ocean floor subsides as it moves away from the ridges. As a result most of the ocean floor (and therefore most of the surface of the earth) obeys a gratifyingly simple relation: its altitude decreases with the square root of its age. In the Atlantic the part of the sea floor that is 100 million years old has moved at a rate of about one centimeter per year to its present position about 1,000 kilometers from the mid-Atlantic ridge, where it formed. As it has cooled it has subsided about two kilometers, giving the ocean there a depth of five kilometers.

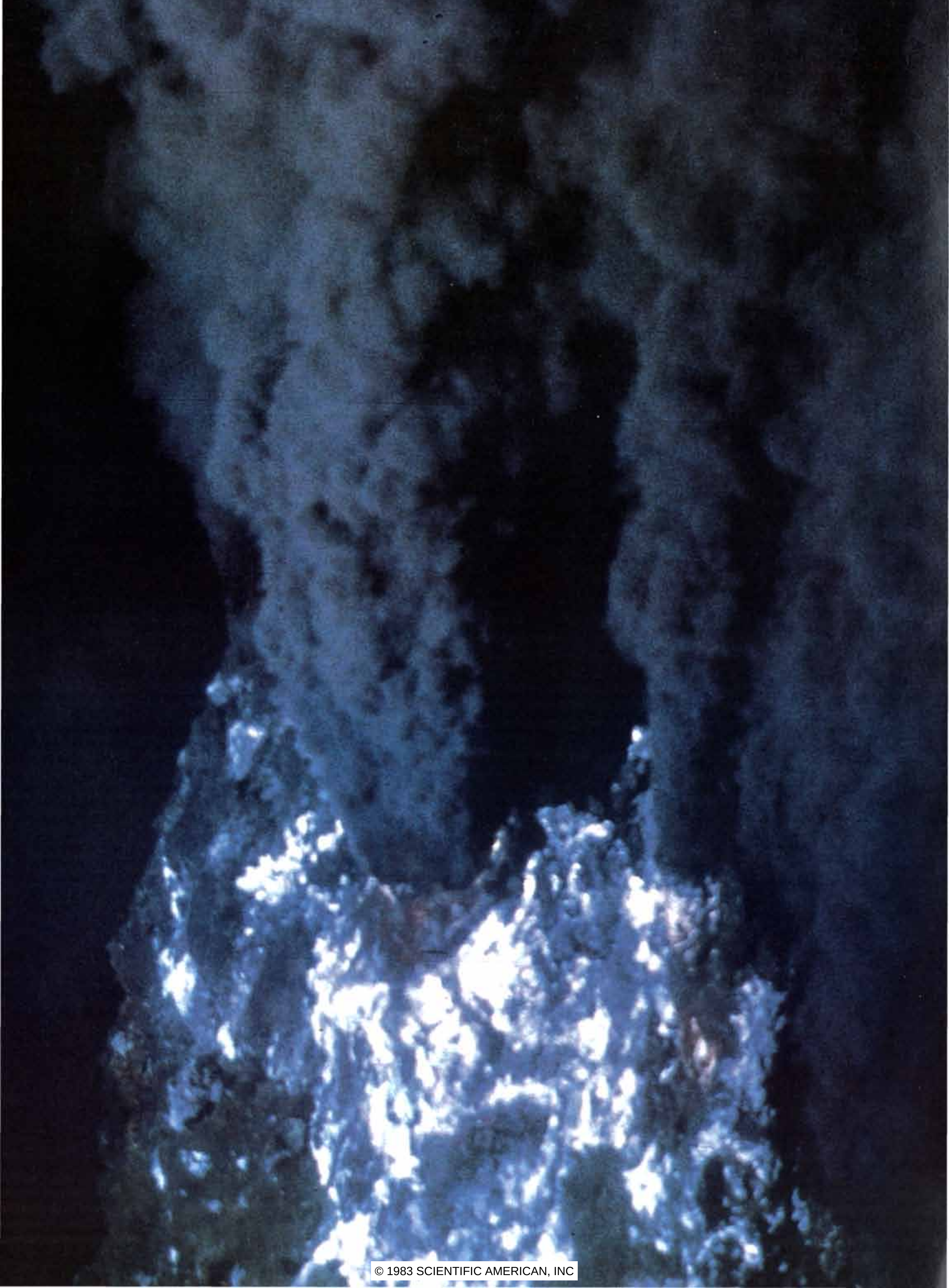
One readily imagines that since the tectonic spreading centers at the midocean ridges are places of volcanic activity, they, like Yellowstone National Park and other areas of volcanic activity on the continents, ought to have hot springs, the sign of hydrothermal activity. Still, one wants to have evidence. The evidence soon came to hand.

In 1966 Kurt G. T. Boström and Melvin N. A. Peterson of the Scripps Institution of Oceanography analyzed sediment samples that had been recovered at sites on each side of the axes of midocean ridges. In this way they found that the newly identified oceanic spreading centers were covered by sediments rich in oxides of iron, manganese and other metals. They proposed that the sediments resulted from hydrothermal reactions of seawater with hot, young oceanic crust. Subsequent work by the Deep Sea Drilling Project, administered by the Scripps Institution, has shown that metalliferous sediment is ubiquitous at the base of the ocean's sediment column, immediately above the volcanic rock of the oceanic crust.

Meanwhile the theory of plate tectonics was pushing the rock of the oceanic crust into the foreground of geologists' attention. The oceanic crust is basalt, a black volcanic rock. Chemically it is an aluminum silicate ($AlSiO_3$) free of quartz (SiO_2) but rich in iron and manganese. (In contrast, granite and rhyolite, the volcanic rocks typical of the continents, are whitish aluminum silicates poor in iron and manganese. In essence they are the slag that has risen to the top of an immense smelting furnace, the earth.)

In the 1960's dredging campaigns

BLACK SMOKER at a midocean ridge is the most dramatic expression of hot springs and hydrothermal activity at the bottom of the ocean. The smoker shown here was photographed from the research submarine *Alvin* at a depth of 2.6 kilometers in the Pacific just south of the entrance to the Gulf of California. It lies on the ridge of the East Pacific Rise. Its diameter is about 1.5 meters; the temperature of the water it vents is 350 degrees Celsius. The water has permeated the oceanic crust, reacted with hot basaltic rock in a zone of volcanic activity under the East Pacific Rise and returned to the ocean floor as a hydrothermal solution: an acidic liquid rich in metals. On its return the solution is cooled by seawater and blackened by a precipitate of iron sulfide particles. The smoker chimney forms from minerals that also precipitate.



were launched to sample outcrops close to ridge axes. The Deep Sea Drilling Project recovered short cores of rock from the bottom of drill holes. The analysis of these samples showed that in general they had been subjected to at least one episode of chemical reaction with seawater after their emplacement as oceanic crust. For example, cracks had developed in the rocks as they cooled to a temperature of less than, say, 500 degrees C., and evidently seawater had permeated the cracks, causing chemical reactions and filling the cracks with precipitated, or secondary, minerals.

This was confirmed by isotopic data. In any chemical reaction involving oxygen the isotopes, or nuclear species, of oxygen are fractionated: they are distributed somewhat inhomogeneously among the reaction products. The fundamental reason is that the vibrational energy of a chemical bond depends on the mass of the atoms bonded together. The degree of fractionation is characteristic of the particular reaction; it also depends on the temperature. The analysis of secondary minerals in the basalt samples from the oceanic spreading centers showed, then, that their composition of oxygen isotopes differed from that of the basalt. In fact, it showed that the minerals had formed by reactions of the basalt with seawater at temperatures as high as 350 degrees C.

Ophiolites

Further evidence of hydrothermal activity at oceanic spreading centers was gathered, surprisingly enough, on the continents. Great slabs of basaltic rock found on top of continental crust on the island of Cyprus, on the northeastern Mediterranean littoral, in Oman on the Arabian Sea, in California, in Tibet and elsewhere were identified as segments of oceanic crust that had been driven onto the continents by plate collisions. Called ophiolites, these formations provided an opportunity to study the mechanisms of formation of the sea floor in detail. In ophiolite terrains one could readily sample and map what had once been oceanic crust, often over hundreds of kilometers horizontally and as much as 10 kilometers vertically. Again the signs of large-scale, high-temperature hydrothermal activity were impressive. In the great ophiolite of Oman, for example, it became plain that seawater had once penetrated to depths in excess of five kilometers and reacted pervasively with the rock at temperatures of about 400 degrees C.

In most ophiolites the basalts are covered by a metalliferous sediment, called umber, that is several meters thick and is identical with the ones Boström and Peterson found on the present-day ocean floor. Moreover, lenticular (lens-shaped) ore bodies containing millions of tons of iron sulfide (FeS_2, known as pyrite or fool's gold) are common in ophiolites. They occupy depressions in the surface of the basalt and are underlain by "pipes" of quartz and ore minerals. Clearly the pipes were once conduits

GIANT CLAMS AT A HOT-SPRING FIELD on the ocean floor just south of the Gulf of California cluster between pillow-shaped mounds of basalt at fissures from which hydrothermal solutions are venting at a temperature of about 17 degrees C. The solutions are

for ascending hydrothermal solutions from which the quartz, the ore minerals and the lenticular bodies precipitated.

In an effort to directly determine the heat balance of the oceanic crust as it evolves, measurements of the conductive loss of heat from the sea floor have been gathered from all parts of the oceans. The conductive loss represents the flow of heat through a motionless medium, in this case the sediment covering the sea floor. In general the conductive loss turns out to decrease toward the midocean ridges. Yet the crust at the ridges must be hotter than the older crust farther away. Hence much of the heat brought to the ridges by ascending magma is being removed by another mechanism, namely convection, the transport of heat by a fluid, in this case seawater. Indeed, Clive R. B. Lister of the University of Washington has calculated that the thermal processes at the ridge axis must be dominated by convection.

The global integration of the anomaly between the measured heat loss and the amount one would expect if the loss were entirely conductive gives a value of about 5×10^{19} calories per year, or about a tenth of the total heat flux from the interior of the earth. This value must be attributed to convection. Suppose the operating temperature of the convective cells is 350 degrees C., in agreement with the isotopic data. Then a volume of seawater equal to the volume of the oceans (1.37×10^{21} liters) must circulate through the ridge axes, and therefore through a temperature regime of 350 degrees C., every eight million years or so. The rate is about half of 1 percent of the rate at which the oceans receive water from the rivers of the earth.

rich in hydrogen sulfide, which nourishes certain bacteria. The bacteria nourish the clams.

Helium Isotopes

Perhaps the strongest evidence that hydrothermal activity is important in the oceans today comes from measurements of the concentration of helium in the oceanic water column. Helium has two isotopes: helium 3 and helium 4. At the time the earth was formed the two isotopes had a certain ratio of abundance with respect to each other. Helium 4, however, is produced inside the earth by the radioactive breakdown of the long-lived isotopes of uranium and thorium. Moreover, both helium 3 and helium 4 are escaping into interplanetary space from the upper atmosphere of the earth, and at the same time both of them are being injected into the atmosphere by the wind of particles that arrives at the earth from the sun.

The ratios of helium 3 with respect to helium 4 from place to place on the earth represent, therefore, the outcome of several processes. Brian Clarke of McMaster University has developed a sensitive mass spectrometer that accurately measures the isotopic composition of helium, and he has employed the device to analyze the helium dissolved in water that trickles from aquifers into wells. The concentration of helium 3 in the water is assumed to have undergone no change since the time the water first entered the aquifer; thus it preserves a record of the equilibrium solubility of helium: the amount that naturally passes into the water from the atmosphere. Any anomaly in the concentration of helium 4 with respect to this equilibrium value can be taken to result from subsurface radioactive decay. In short, the spectrometer is an aid to uranium prospecting.

With that much established, Clarke, in collaboration with Harmon Craig of Scripps, attempted to estimate the leakage of helium 4 from the mantle into the ocean. As before, they assumed that helium 3 would serve as an index. What they found instead was remarkable. At the surface of the Pacific the helium dissolved in the sea was in fact in equilibrium with the helium in the atmosphere. At greater depths the amount of helium exceeded the equilibrium value, and when the equilibrium value was subtracted from the measured concentration, the amount of helium 3 in the excess was anomalously large. The anomaly attained its maximum at 2,600 meters, the average depth of the Pacific ridge axis. There the ratio of helium 3 to helium 4 in the excess helium was about eight times the atmospheric value. Evidently helium 3 is retained in the mantle and is being released into the ocean at the ridge axis. Does the helium simply escape from gas bubbles near the surface of cooling magma or is it liberated in the course of hydrothermal activity that pervasively reworks the rock? The scale of the anomaly strongly indicated the latter.

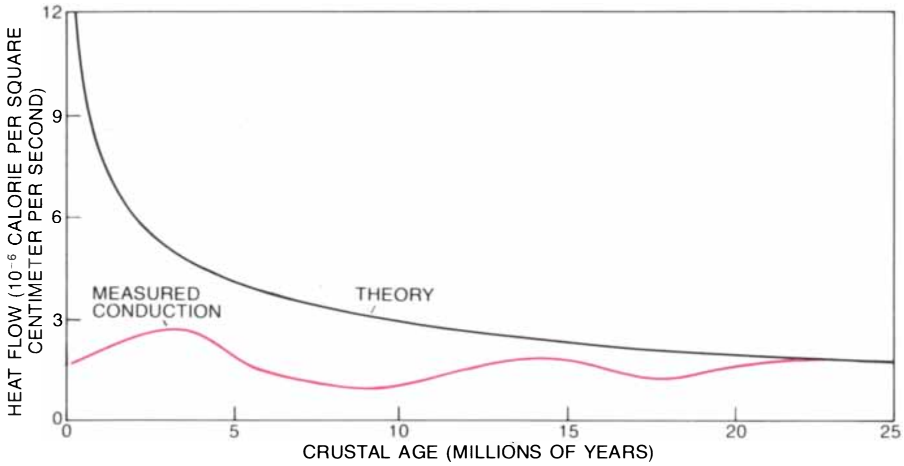
The Composition of Seawater

Precisely how does seawater react with basalt at sea-floor pressures of almost two tons per square inch and at temperatures of several hundred degrees C.? The first answers emerged in the mid-1970's when James L. Bischoff and Frank W. Dickson reproduced those conditions in their laboratory at Stanford University. The results were dramatic. Magnesium ions in the seawater combined with silicate in the basalt to form a hydroxysilicate, $Mg(OH)SiO_3$, which is insoluble in water. The hydroxyl (OH) groups were contributed by the water, leaving hydrogen ions (H^+) behind; thus the seawater solution became highly acidic. Hydrogen ions then took the place of calcium and potassium at sites in the basalt's crystal lattice, and so the lattice recrystallized.

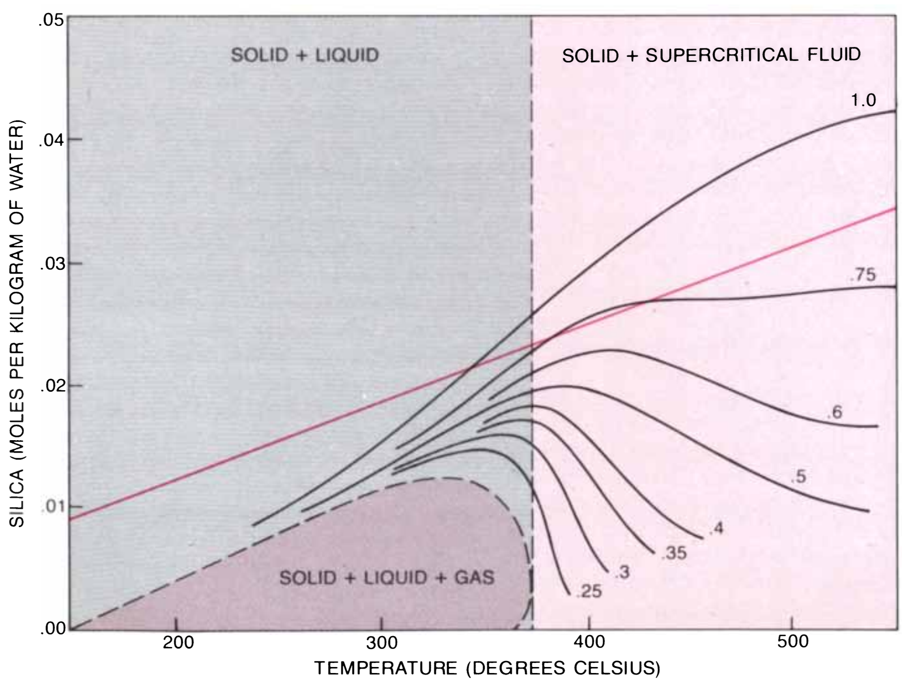
In essence the reactions were an exchange of hydrogen ions for other cations, or positive ions, notably those of calcium and potassium. The liberated calcium could now react with sulfate (SO_4) in the seawater to form calcium sulfate ($CaSO_4$), which precipitates as the mineral anhydrite. Sulfate could also react with iron in the basalt to form iron sulfide, that is, deposits of pyrite. The result of these two reactions was the complete removal of sulfate from the seawater. Finally, the degradation of the basalt and the acidity of the seawater solution allowed the release of silica (SiO_2 dissolved in water).

Might the chemical composition of the oceans come as much from these hydrothermal reactions as it does from the products of weathering on the continents? The possibility was particularly exciting because in the mid-1970's the understanding of the processes that control the composition of seawater were at an impasse. In the 19th century, before the boom in organic chemistry that followed the discovery of synthetic dyes, many prominent chemists had undertaken analyses of inorganic natural substances.

One important motivation was the search for unknown chemical elements to fill the gaps in the periodic table of the elements, which was then being developed. The chief strategy was the analysis of ore minerals. R. W. Bunsen, for example, analyzed hundreds of ores. He became interested in magmatic processes and traveled to Iceland to witness volcanic eruptions. Other investigators directed their attention to the study of natural waters. Thus J. H. van't Hoff laid the foundation of experimental geochemistry by studying the sequence of precipitates that form as seawater evaporates.



FLOW OF HEAT from the ocean floor was an early sign that hydrothermal activity is important at midocean ridges. At each ridge, plates of oceanic crust diverge and new crust rises into place; hence the ridges are notably hot. Yet the measured conduction of heat through the crust at three Atlantic ridges (*color*) proved to be markedly less than the amount to be expected if the flow were entirely by conduction (*black*). The difference must be carried by convection, that is, water passing through the ocean floor. The horizontal axis of the chart shows the age of the oceanic crust, a correlate of its distance from the midocean ridge where it was emplaced.



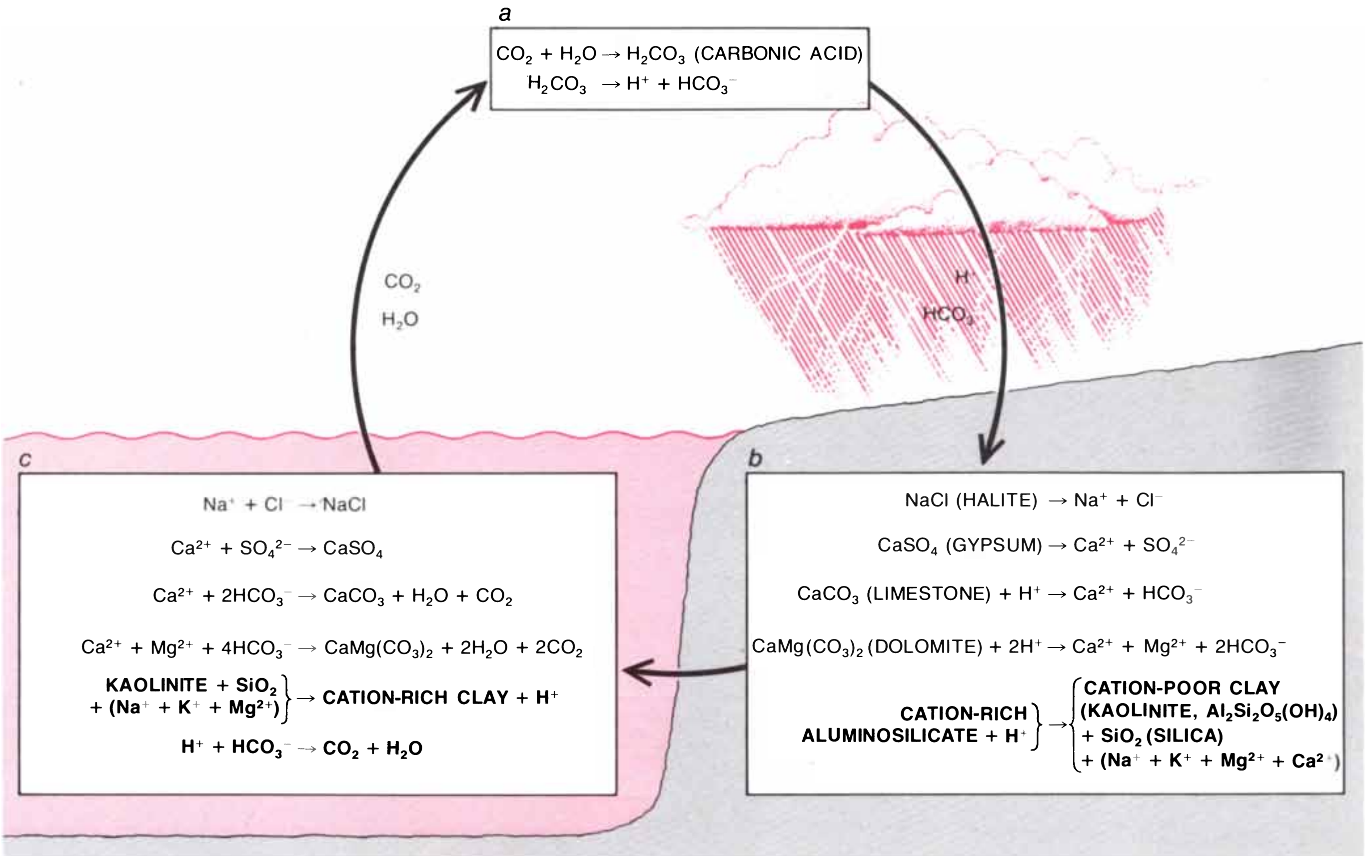
SOLUBILITY OF QUARTZ in vent water sampled in 1977 at ocean-floor hot-spring fields near the Galápagos Islands suggested even before black smokers were discovered (two years later) that pure hydrothermal solutions rising toward the ocean floor might turn out to have a temperature of about 350 degrees C. The Galápagos samples were no hotter than 19 degrees; still, the concentration of silica (dissolved quartz) was greatest in the hottest samples. The extrapolation of the trend (*color*) intersected curves representing water saturated with quartz (*black*) at points well above 300 degrees. (Pressures are labeled in kilobars.) Evidently the Galápagos samples represented hydrothermal solutions diluted with ordinary seawater. The chart shows the regimes in which a quartz-water mixture is stable as (1) solid plus liquid plus gas, (2) solid plus liquid and (3) solid plus supercritical fluid (a phase that is neither liquid nor gas).

From all these efforts it was plain by the end of the century that seawater could not be produced by the partial evaporation of river water. At the end of that route lie only closed-basin lakes such as the Dead Sea and Great Salt Lake, which are highly alkaline compared with the oceans. Then came a long hiatus. It lasted until the late 1950's, when Lars Gunnar Sillén of the Royal Institute of Technology in Stockholm undertook to clarify matters. Sillén asked: What controls the *p*H, or hydrogen-ion concentration, of the oceans? Why is it consistently 7.5 to 8, or quite close to acid-alkaline neutrality? Subsequently Frederick T. Mackenzie and Robert M. Garrels of Northwestern University translated Sillén's rather abstract answer into geological terms.

Consider the weathering of volcanic rocks on the continents. The rain beating down on the rocks is an acid solution because carbon dioxide from the atmosphere dissolves in water to make carbonic acid (H_2CO_3). For their part the rocks are an aluminum silicate lattice in which cations are embedded. The acid reacts with the lattice. Specifically, the acid gives up hydrogen ions, leaving bicarbonate (HCO_3^-) behind; thus the rainwater becomes alkaline. The hydrogen ions take the place of cations in the lattice, freeing them into solution. The lattice is therefore disrupted; it is transformed into highly disordered clay minerals such as kaolinite ($Al_2Si_2O_5(OH)_4$), which are easily eroded.

Estimates of the global rate of these reactions can be derived from the rate at which their products are carried toward the oceans by the rivers. The result requires that an amount of carbon dioxide equal to its abundance in the atmosphere be consumed in about 4,000 years. Clearly there must be a reaction that converts bicarbonate back into carbon dioxide. Mackenzie and Garrels noted that the ratio of cations to hydrogen ions is about 1.2×10^3 in average continental waters. It is 6×10^7, four orders of magnitude greater, in the oceans. Thus when aluminum silicates rich in hydrogen ions are transported into the oceans (where hydrogen ions are scarce and cations are plentiful), the reactions that made them should reverse. Clays rich in cations should form; hydrogen ions should be released and then should combine with bicarbonate to make water and carbon dioxide.

The implications were profound. The gross composition of the continental crust has been constant for at least the past 2.5 billion years; hence the composition of the oceans would also have been unchanging over that vast period. Studies were undertaken in an attempt to establish that the crucial reactions re-



CHEMICAL BALANCE between the atmosphere, the continents and the oceans was hypothesized in the 1950's by Lars Gunnar Sillén of the Royal Institute of Technology in Stockholm and worked out in detail by Frederick T. Mackenzie and Robert M. Garrels of Northwestern University. In the atmosphere (*a*) carbon dioxide is dissolved in water vapor as carbonic acid. The water rains or snows on the continents (*b*), where hydrogen ions from the acid solution leach cations (positive ions) such as calcium, magnesium, sodium and potassium out of the continental rocks. Rivers carry the results of the continental weathering to the oceans (*c*). In the oceans, which are poor in hydrogen ions, the reactions reverse, creating sedimentary rock rich in cations and restoring carbon dioxide to the atmosphere.

constituting clay minerals were in fact happening in the oceans. The findings were disappointing. Even the most optimistic interpretation of the data could not approach the scale necessary to close the balance between the land and the sea. Here, then, was the impasse mentioned above. The preliminary calculations of the chemical fluxes to be expected from ridge-axis hydrothermal activity pointed to a way out.

First Explorations

Of course, being hopeful that hot springs are common at the oceanic spreading centers and actually finding them under at least 2.5 kilometers of water are very different things. When the first attempts were made in the early 1970's, knowledge of the sea floor and the capacity to explore it were equally rudimentary. The sonar devices then in service were adequate for flat terrain. At the rugged topography of the ridge axes, however, they recorded a jumble of echoes. As a result the equipment towed close to the bottom in efforts at exploration was in effect pulled through elaborate obstacle courses in the dark. Instruments were damaged and lost at an alarming rate. Nevertheless, the most sophisticated device, Deep Tow, operated by the Scripps Institution, made significant findings.

Deep Tow, a vehicle towed at the end of a telemetry cable from a ship on the surface, carried television cameras, sonar, pressure sensors and probes to measure seawater temperature and electrical conductivity (a correlate of salt content). Over several years it detected temperature anomalies at a number of places on the spreading centers in the eastern tropical Pacific. In one case Deep Tow carried water samplers, and Ray F. Weiss of Scripps managed to catch a sample from the midst of an anomaly. The temperature of the water differed by less than .1 degree C. from the temperature of the ambient water: two degrees. Still, the chemical data, including the concentration of helium 3, showed without doubt that the water was of hydrothermal origin. Presumably Deep Tow had "flown" into a plume 15 or 20 meters above a submarine hot spring.

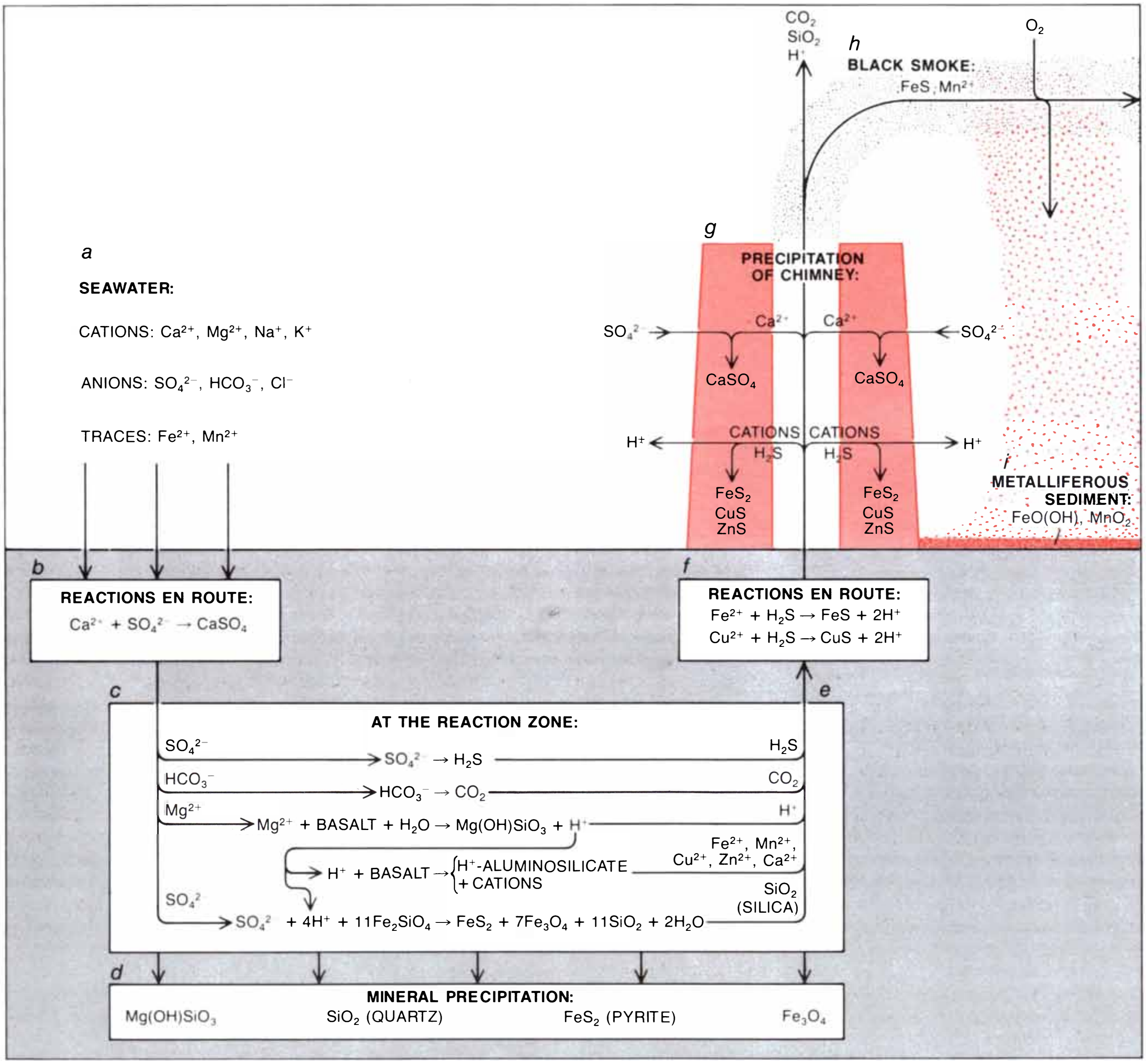
In the mid-1970's changes began to come rapidly. For one thing, the U.S. Navy made available to researchers the techniques it had developed for mapping the ocean floor. Thus high-precision deep-sea navigation systems came into routine service. Such systems depend on the measurement of the time that elapses between the broadcast of an acoustic pulse (an "interrogation pulse") from an undersea vehicle and the reception of acoustic "replies" from an array of transponders moored to the bottom. If the relative positions of the transponders have been determined by sonar from a ship at the surface, the undersea vehicle can be navigated quite readily. Indeed, positional accuracies of a few meters are achievable.

Second, a large-field survey camera was developed. The available cameras had been flimsy, lightweight devices towed far behind a ship, so that they were difficult to control. What was needed was a massive camera vehicle. Angus, developed at the Woods Hole Oceanographic Institution, was the result. It is a 1.5 ton "gorilla cage" in which are mounted color cameras, stroboscopic lamps, power supplies, sonar and an acoustic navigation transponder. In its current configuration the vehicle is towed from a ship at a rate of about four kilometers per hour at elevations of about 20 meters above the sea floor. Because of its mass, it is never more than 75 meters from being directly under the ship; hence the ship's sonar readings can help to keep the vehicle from harm. Normally a tow of Angus lasts for 18 hours. A color photograph is made every 10 seconds. The vehicle is brought up, the film is developed and intriguing features are noted. Each photograph bears a record of when it was made; from this datum the feature can be located exactly.

The third crucial technology is the research submarine. The submarines that can work at ridge-axis depths are small. The largest of them, *Alvin*, operated by Woods Hole, weighs 16.5 tons and carries two investigators and a pilot in a titanium pressure-resistant sphere two meters in diameter. It is severely limited in power. Thus it is slow: its maximum speed over the bottom is about four kilometers per hour. Moreover, its lights can illuminate the water to a distance of only about 15 meters. It is an inefficient tool for exploration. Its proper role, however, is in visiting sites already identified (say in photographs made by Angus) as deserving study. *Alvin* is directed to such targets by means of the same navigation transponders that guide Angus. Routinely one can land *Alvin* on the sea floor within a few tens of meters of the designated target.

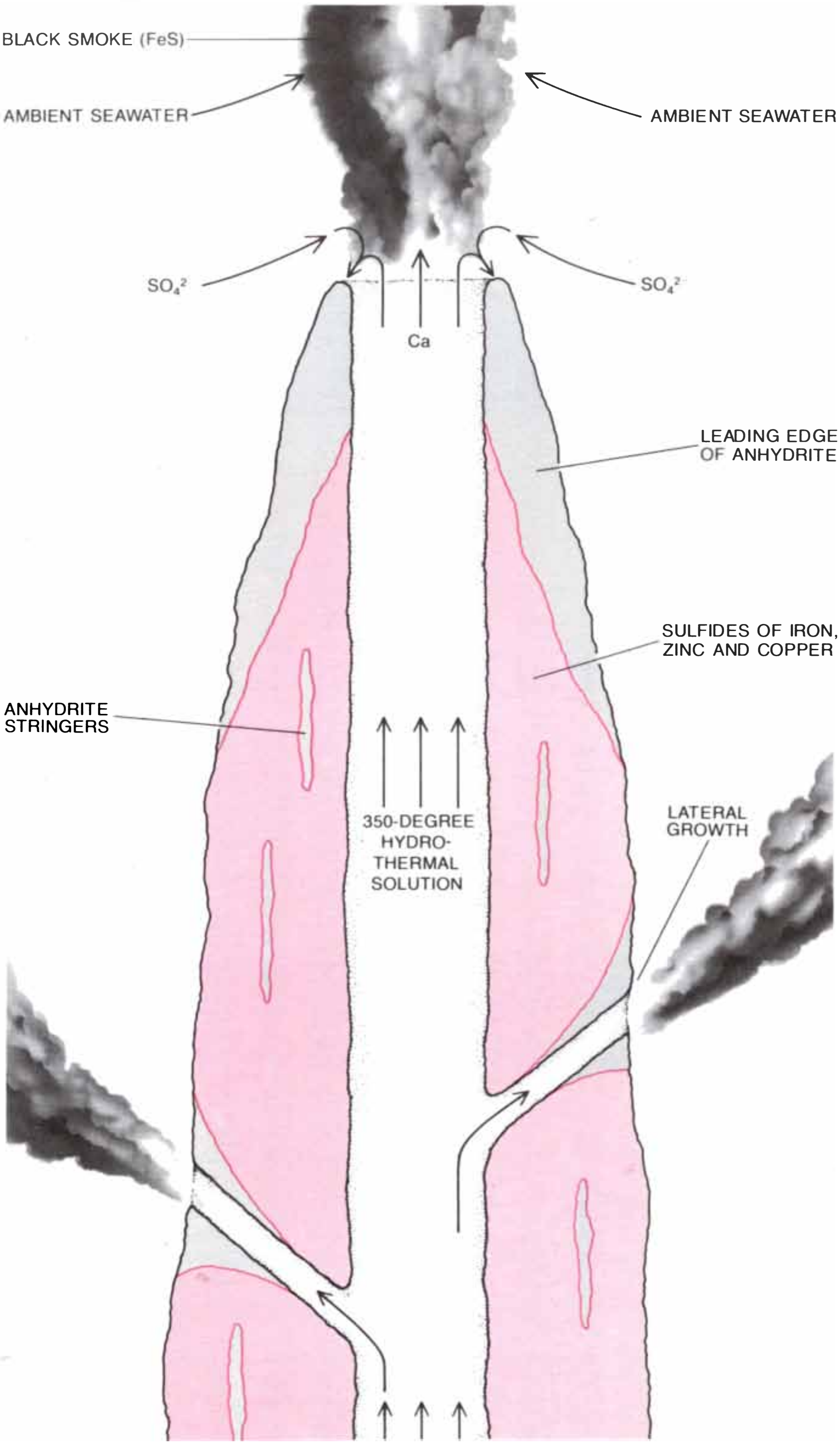
A Fabulous Scene

This elaborate and costly three-part system was first deployed in the spring of 1977 at a spreading ridge in the Pacific 280 kilometers northeast of the Galápagos Islands. The research vessel



HYDROTHERMAL REACTIONS at the midocean ridges regenerate carbon dioxide to a far greater extent than the low-temperature oceanic process Sillén envisioned. Seawater (*a*) percolates into the oceanic crust. As it descends some of its content of ions may precipitate (*b*) as calcium sulfate. The remaining ions reach a zone several kilometers under the ocean floor. There they react with hot basaltic rock (*c*) to produce mineral precipitates (*d*) and a hot, acidic, metal-rich hydrothermal solution (*e*) that rises back toward the ocean floor. The rising solution may encounter cold seawater and cool, thus precipitating metal sulfides (*f*). Then, at the ocean floor, the solution mixes with sulfate from seawater to precipitate chimneys consisting of metal sulfides and calcium sulfate. Above the chimneys the iron still in solution forms black iron sulfide smoke (*h*). Manganese stays in solution. Eventually the iron and the manganese are oxidized and rain onto the ocean floor as a metalliferous sediment that is ubiquitous atop the oceanic crust (*i*). Meanwhile the carbon dioxide emerging from the vents is mixing into the ocean. Eventually it reaches the surface, enters the atmosphere and closes the carbon dioxide cycle.

Knorr, out of Woods Hole, arrived first. It laid down a network of transponders and surveyed their positions. Then Angus began its runs, under the direction of Robert D. Ballard of Woods Hole. By the time *Alvin* arrived a few days later on its tiny mother ship *Lulu,* several targets had been identified. At each one the Angus photographs showed a few large white clams on a black basaltic field. *Alvin* was directed to one of the targets. The investigators aboard were John B. Corliss of Oregon State University and one of us (Edmond).



BLACK SMOKER ACCRETES from calcium sulfate, which precipitates out of a hydrothermal solution as a "leading edge" of the whitish mineral anhydrite. As the leading edge advances it becomes exposed to hot, undiluted solution. As it redissolves, however, it acts as a template for metal sulfides. Anhydrite is preserved in places where sulfides engulf it rapidly.

Alvin sinks passively in the ocean's water column at a rate of 30 to 35 meters per minute. Thus it settled to a depth of 2,500 meters in one and a half hours. At that point, 100 meters from the bottom, it jettisoned a set of weights and gained neutral trim. We moved downward until we hovered above the sea floor, which proved to be a gentle slope. For about half an hour we cruised about, trying to find the target. Each of us had a Plexiglas porthole to look through. Then we stopped to collect some rocks. As our pilot employed the submarine's mechanical arm to wrestle with a basalt pillow (a volcanic deposit that results from the slow extrusion of lava on the sea floor when the surface of the lava chills, hardens, cracks and heals so that the lava inflates itself into the shape of a pillow) a couple of large purple sea anemones engaged our attention. Only when our gaze finally wandered away from them did we realize that the water within the range of our lights was shimmering, like the air above a hot pavement. The hastily measured temperature of the water was five degrees above the ambient water temperature (2.05 degrees C.). With all thoughts of rocks forgotten, we captured a sample of the water and then continued on our course upslope. Here we came on a fabulous scene.

The typical basaltic terrain at the ridge axis is bleak indeed. Monotonous fields of brown pillows are cut by faults and fissures. One must examine several square meters to find a single organism. Yet here was an oasis. Reefs of mussels and fields of giant clams were bathed in the shimmering water, along with crabs, anemones and large pink fish. The remaining five hours of "bottom time" passed in something close to frenzy. We recorded the temperature, conductivity, *p*H and oxygen content of the water; we made photographs; we sampled the water; we made sure that a representative selection of organisms was collected, all under the growing pressure of steadily decreasing voltages for our equipment.

Fortunately the equipment worked flawlessly. It soon became apparent that we had come on a hot-spring field. Warm water streamed from every orifice and crack in the sea floor over a circular area about 100 meters in diameter. The temperature of the water was highly variable, but the maximum was about 17 degrees C. The organisms were quite selective. They choked the warmest vents. In some cases mussel reefs actually channeled the flow of water, forming conduits themselves.

We worked until the "scientific power" ran out. Then a second set of weights was jettisoned and *Alvin* left the bottom. The equipment was shut down. Soon we began to get cold. After eight hours of crouching in the two-meter sphere, with almost every movement requiring the

cooperation of one's equally cramped colleague, we spent the 90-minute ascent trying to keep warm, too tired even to watch the bioluminescent organisms disturbed by the submarine's passage. The approach to the surface, as always, was like a rebirth. At a depth of about 200 meters a faint green glow began to pervade the water. Soon all was bright. The pilot vented the ballast tanks, surrounding *Alvin* with a column of bubbles. There was a soft bump as we surfaced; then the lazy roll in the swell. Shortly divers were around the submarine attaching recovery lines. *Alvin* was maneuvered onto its recovery cradle. Then came the agony of standing up and clambering out. By the time we had staggered onto *Lulu,* everyone shared our excitement. *Alvin's* sample basket was loaded with clams and mussels. Crabs, which had got inside the fiberglass skin of the submarine, began to fall out on deck.

Living on Hydrogen Sulfide

Now the real effort began. The water samplers were disassembled. The water samples were transferred to the *Knorr* and analyzed through the night. The Angus photographs were scrutinized for more targets. The water samplers were cleaned and reassembled. At 6:00 A.M. the next day *Alvin* was being readied; at 8:45 it descended again. It made a total of 15 dives.

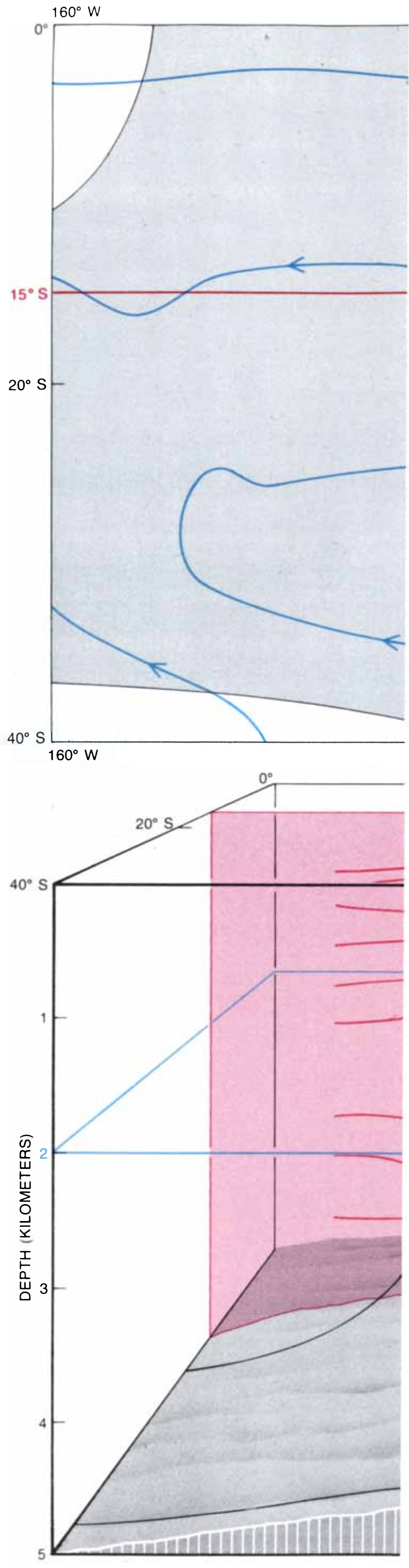
As the analytical work progressed it became clear that all the hot-spring waters had a high concentration of hydrogen sulfide (H_2S). Herein lay the explanation of the ocean-floor oases. Bacteria that derive their energy from the oxidation of hydrogen sulfide are common in many ecological systems. At the oases, however, they had to be the primary producers at the base of the ecological pyramid. Fundamentally the energy driving the newly discovered ecosystem flows not from the sun, as in photosynthesis, the source of energy for all other ecosystems, but from the radioactive breakdown of long-lived isotopes of uranium, thorium and potassium inside the earth. The breakdown releases heat; the heat produces magma, which becomes new oceanic crust. Seawater percolates into the crust and reacts with the crustal rock at high temperature and pressure.

Two of the reactions are crucial for the oases. Sulfate from the seawater reacts with iron in the rock to produce hydrogen sulfide and iron oxides; moreover, sulfide minerals in the rock are dissolved. Thus the hot water rising back toward the ocean floor is heavily charged with sulfide. At vents on the sea floor the hydrogen sulfide is absorbed by bacteria that also take up oxygen of photosynthetic origin dissolved in the ambient water. The sulfide and the oxygen combine, so that the sulfide is turned back into sulfate. The reaction releases the energy that drives the bacteria's metabolism. The bacteria in turn nourish other species. The diversity of the species was an immediate indication that hot springs must be common on the ridge axes of the world. How else could a unique and highly evolved fauna be developed and sustained?

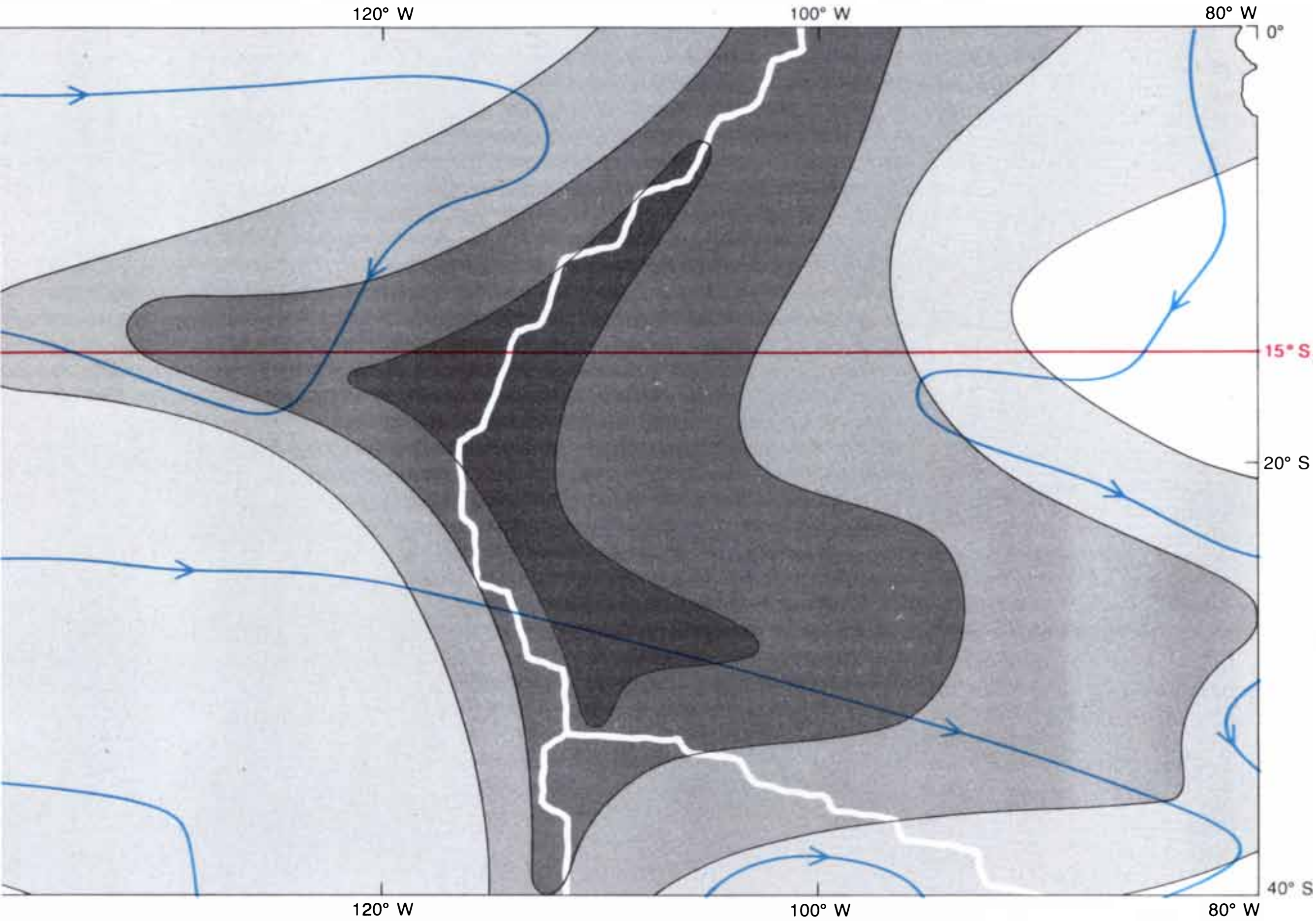
Our own laboratory at the Massachusetts Institute of Technology had been assigned to analyze in detail the water *Alvin* had sampled at the vents. Plainly the chemistry of the water was the result of a complex set of reactions between seawater and basalt that proceed in an inaccessible reaction zone, perhaps several kilometers under the ocean floor. Hence the best way to go about unraveling the reactions would be to see how the water samples were enriched or depleted in the various chemical elements with respect to basalt and with respect to ordinary seawater. In the end we measured the concentration of 35 elements, probably the most ever measured in an effort of this type. As we progressed the chemical dynamics began to unfold.

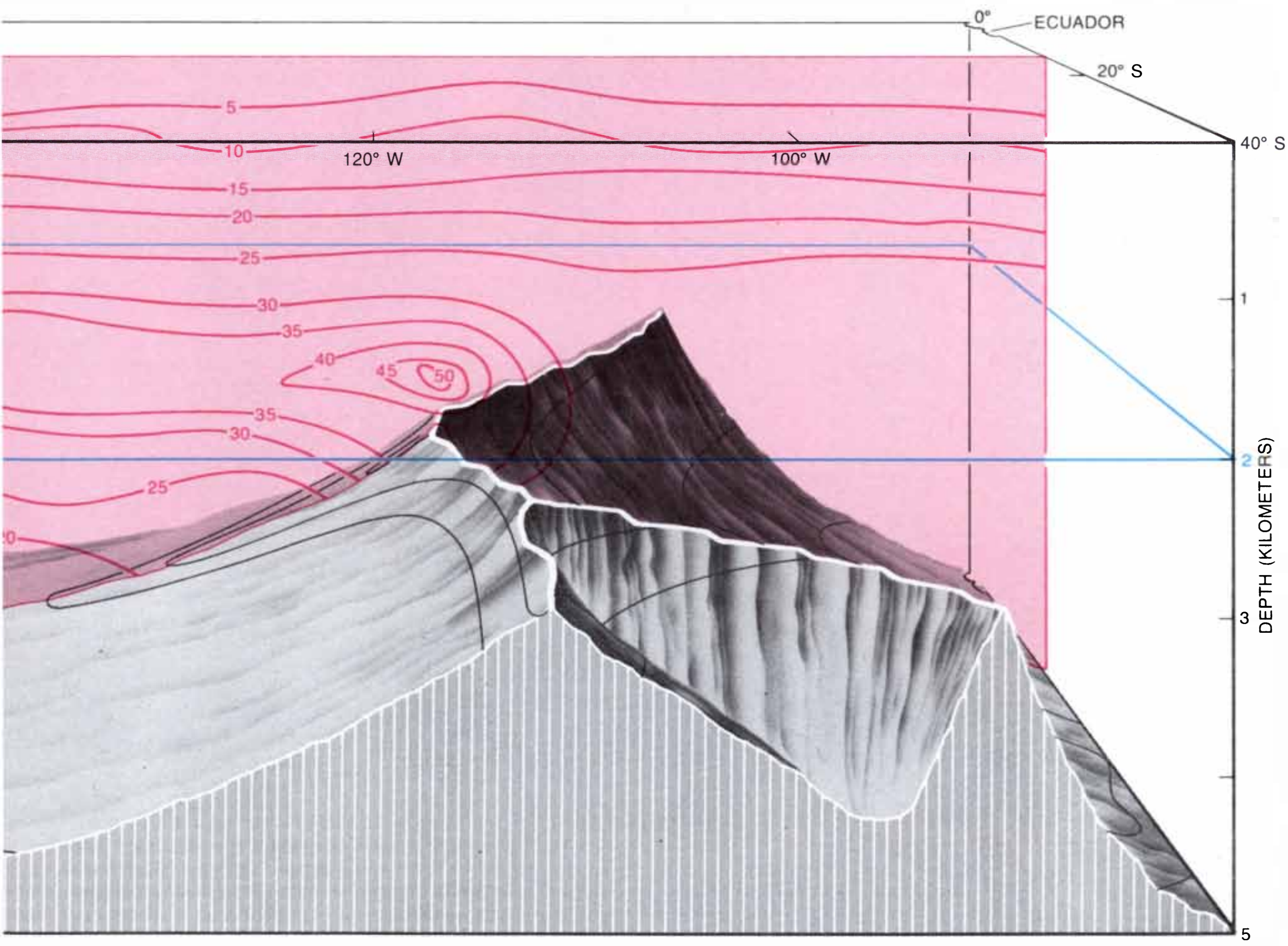
The Vent Water's Properties

Magnesium and sulfate, which are constituents of ordinary seawater, showed a uniform decrease in concentration with increasing temperature



DISTRIBUTION OF METALLIFEROUS SEDIMENT on the floor of the South Pacific, as mapped by Kurt G. T. Boström of the Scripps Institution of Oceanography, has features that make it asymmetric with respect to the ridge axis from which the metals issue in a hydrothermal solution. The drawing at the top displays the South Pacific from above. At 15 degrees south latitude a long, narrow salient of sediment rich in oxides of iron and manganese projects westward from the axis; it is indicated in shades of gray. At 30 degrees south latitude a broader, shorter salient projects eastward. Blue lines mark ocean currents at a depth of 2,000 meters (about 600 meters above the ridge axis) as they have been deduced by Joseph L. Reid of the Scripps Institution. The currents have "blown" the oxides away from the axis; thus the currents account for the salients. The drawing at the bottom displays the South Pacific from the south. Red lines mark concentrations of the isotope helium 3 in a plume extending westward across the Pacific at 15 degrees south latitude. The pattern was measured by John E. Lupton and Harmon Craig, also of Scripps. Numbers indicate the concentration of helium 3 as a percent in excess of its normal saturation value in the sea. Evidently helium is freed from the oceanic crust by hydrothermal reactions. Because it is inert, it undergoes no change as it is vented at the axis; hence it labels the ocean water from which metalliferous sediment precipitates.


120° W
100° W
80° W
0°
15° S
20° S
40° S
120° W
100° W
80° W


0°
ECUADOR
20° S
40° S
120° W
100° W
5
10
15
20
25
30
35
40
45
50
35
30
25
1
2
3
5
DEPTH (KILOMETERS)

(that is, the temperature of each water sample as *Alvin* had trapped it). Clearly each sample represented water that had issued from vents and then been diluted to some degree by the ambient seawater. The extrapolation of the concentration-temperature trends to a concentration of zero, on the hypothesis that pure vent water might prove to have no magnesium and no sulfate in it, yielded a temperature of about 350 degrees C. The extrapolation was admittedly extravagant: the hottest water we had encountered had a temperature of only 19 degrees. On the other hand, the concentration of silica (quartz dissolved in water) increased with increasing temperature. The extrapolation of the trend to its point of intersection with curves representing the solubility of quartz in water at various pressures and temperatures also yielded a temperature of about 350 degrees.

Iron, copper, nickel and cadmium, which combine with hydrogen sulfide in the vent water to form sulfides or oxides that precipitate out of water, all showed marked decreases with increasing temperature. Their trends extrapolated to zero at a range of temperatures from 30 to 35 degrees C. That range could be interpreted as the temperature of a reservoir of water being tapped by the hot springs. We surmised that the water ascending from a regime of high-temperature reactions entrained "groundwater" whose temperature and composition were close to those of the ambient seawater overlying the ridge axis. The high-temperature solution, or "end member," was cooled by this groundwater, and its acidity was diluted, causing the precipitation of the sulfide- and oxide-forming elements in the "plumbing" of the basaltic crust. The precipitation proceeded to such an extent that the sulfide- and oxide-forming elements were stripped even from the groundwater. They would then have been contributed to our samples only by the ambient water that diluted the samples. Evidence of the high-temperature reactions was preserved, however, in the magnesium, sulfate and silica data.

Helium 3 increased strongly with temperature, at a ratio of 2.2×10^{-17} moles (about 13 million atoms) for every calorie of heat. In the helium that remained when the equilibrium amount was subtracted from the measured concentration the ratio of helium 3 to helium 4 was eight times what it is in the atmosphere. In the wake of their discovery of the helium-3 anomaly deep in the Pacific, Craig, Clarke and their colleagues had calculated that the amount of helium 3 required to maintain the anomaly worldwide against the loss of helium to the atmosphere as the oceans overturn is some 1,100 moles per year. If all the helium 3 entering the oceans enters by way of ridge-axis hot springs, and if the ratio of helium 3 to heat is everywhere the same as the one measured at the Galápagos ridge axis, the worldwide transport of heat at the ridge-axis hot springs is readily calculated. It is 5×10^{19} calories per year, which places it in the middle of the range the geophysicists had calculated.

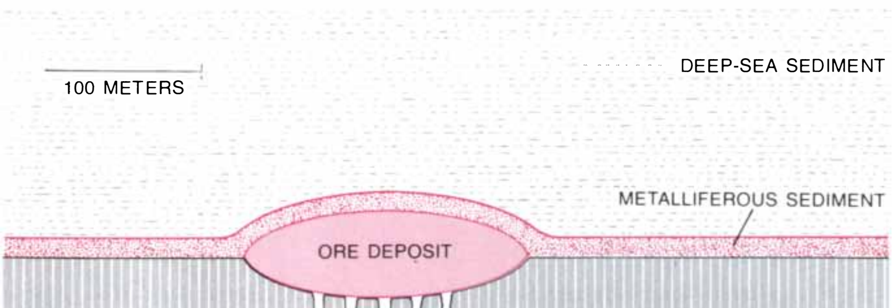
One could feel justified, therefore, in supposing that anomalies in the concentration of other elements in our samples would serve (as proportions of the helium-3 anomaly) to calculate the total fluxes of these elements from submarine hot springs into the oceans. The results are impressive, particularly when they are compared with the fluxes from continental weathering. Thus the ridge axes consume most of the magnesium and most of the sulfate that rivers introduce into the sea. Conversely, they release between five and 10 times more lithium and rubidium, and between a third and a half as much potassium, calcium, barium and silica. Their release of manganese is sufficient to account for all the accumulation of that element in the metalliferous sediments and nodules carpeting the ocean. Finally, the ridge axes convert back into carbon dioxide most of the bicarbonate produced by continental weathering. The low-temperature oceanic reactions Sillén envisioned are therefore replaced by a high-temperature hydrothermal process.

Black Smokers

It appeared from the analyses that conceptions of the processes controlling seawater chemistry over geologic time would need to be substantially modified. It would be well, of course, to discover hot end-member solutions emerging from the sea floor undiluted. The pervasive entrainment of groundwater at the Galápagos ridge axis made that prospect seem most doubtful. It was with surprise and excitement, therefore, that we heard news from Jean Francheteau of the Institut de Physique du Globe. The French research submarine *Cyana* had found large sulfide deposits on the ridge axis at the crest of the East Pacific Rise just south of the mouth of the Gulf of California at 21 degrees north latitude. The site gave no sign of current hydrothermal activity. Nevertheless, the deposits could have formed only by the precipitation of sulfides from a high-temperature solution.

We ourselves returned to the Galápagos ridge axis in the spring of 1979. Many more vent fields were found, but the water temperature was no greater than 23 degrees C. *Alvin* then left to take part in the work near the Gulf of California. Angus, already there, had photographed fields of hot springs, and in the midst of them *Alvin* found what we had thought would never be found. Great jets of black water spewed from chimneys of sulfide minerals that had built themselves up to heights of several meters. The temperature of the vent water exceeded 300 degrees C. (At the pressure of sea level, water at that temperature would boil explosively, as it does in geysers such as Old Faithful in Yellowstone. At the pressure of these depths, however, it remains liquid.) A scramble ensued to design and build water samplers that would function at temperatures exceeding 300 degrees. The National Science Foundation allocated additional time for *Alvin* to dive to the "black smokers." John A. Archuleta of the Los Alamos National Laboratory lent us equipment.

We dived in November, 1979. Navigation transponders left at the site guided us to the vents. We approached a black smoker issuing from the sea floor amid large blocks of sulfide ore. Maneu-



OPHIOLITE, a segment of oceanic crust emplaced on top of a continent, includes layers of sediment identical with the ones that carpet the ocean floor. It also contains sulfide ore bodies that are taken to have resulted on the ocean floor from the hydrothermal activity of a multitude of black smokers. Often the ore bodies occupy depressions in the top of a layer of basalt.

vering in a vigorous bottom current, the pilot inserted a newly designed temperature probe into the throat of the smoker, only 15 centimeters in diameter. The reading stabilized within a few tenths of a degree of 350 degrees. The predictive power of chemistry was vindicated! All the vents we sampled differed from a temperature of 350 degrees by no more than a few degrees.

The 350-degree solutions leave the vents as clear, homogeneous fluids. They contain 100 parts per million of iron and a few parts per million of zinc, copper and nickel. That represents enrichment by factors as great as 10^8; the concentration of iron, zinc, copper and nickel in ordinary seawater is measured in parts per trillion. In addition the solutions contain 210 parts per million of hydrogen sulfide; ordinary ocean water contains none. The concentration of silica is 1,290 parts per million, exactly what one would predict after examining the Galápagos data. In other words, the solutions are saturated with quartz. Conversely, the solutions contain no magnesium and no sulfate. That too is exactly what one would predict.

At the mouth of a vent the solutions quickly mix with the cold ambient seawater. The result is the black "smoke." It is a suspension of fine particles of iron sulfide that precipitate out of the solutions. The chimneys appear to grow by the accretion of a leading edge of calcium sulfate. The calcium is contributed by the hydrothermal solutions and the sulfate by seawater. As the leading edge advances it becomes exposed to the undiluted vent waters. Hence it redissolves, to be replaced by precipitated sulfide minerals. Vent water commonly breaks through leaks in the chimney, promoting lateral growth. The result is a complex multimineral deposit.

Exploiting the Ores

The vent solutions themselves would not be worth the effort to mine them. The recovery of a ton of zinc, for example, would call for the processing of about six million tons of solution, the amount that flows out of a typical black smoker over a period of several months. A ton of zinc is worth about $900. Plainly it is better to let nature take its course and exploit the resulting ore bodies.

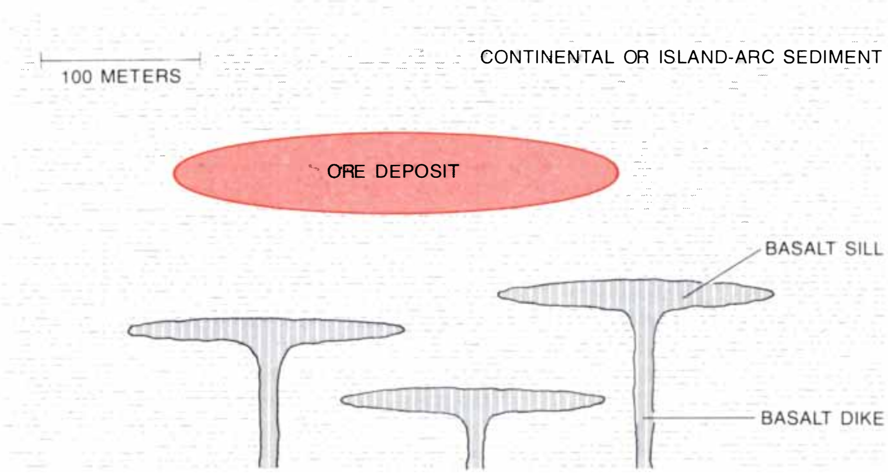
First consider the black smokers. The formation of a massive sulfide deposit such as the deposits in ophiolites, which contain millions of tons of ore, would seem to require a forest of smokers. Perhaps the smokers originally occupy a depression in the sea floor; particles precipitating from the smoke might then fill the depression. In any event the hydrothermal deposit accumulating around the smokers will be reworked by hot solutions rising through it in their ascent toward the ocean floor. The deposit will be stripped of its minor elements (copper, zinc, cobalt and nickel), leaving a matrix of large crystals of pure pyrite. The minor elements will be redeposited above the pyrite, in chimneys at the ocean floor. Meanwhile the leakage of cold seawater into the accumulating deposit will favor the in situ precipitation of fresh deposits from the ascending hydrothermal solutions.

This sequence accounts for the general features of the deposits that man has exploited in ophiolites. The main body of such a deposit is composed of pure, coarse-grained pyrite. The layer above the pyrite, which has oxidized to ocher, is enriched in the minor elements. Moreover, Alexander Malahoff of the National Oceanographic and Atmospheric Administration has recently discovered a forest of black smokers. About 30 kilometers from the Galápagos fields he observed from *Alvin* a profusion of extinct chimneys along a 600-meter length of the ridge axis. They were on top of a sulfide body several tens of meters high. Given the large amount of heat the chimneys must have released when they were active it is questionable whether they could have been approached at such a time by a research submarine. The Plexiglas portholes of *Alvin* soften at 86 degrees C.

It must be said that the deposition of ore at the site of black smokers themselves is extremely inefficient. An overwhelming fraction of the metals dissolved in the hydrothermal solutions become particles of the smoke that leaves the chimneys; this fraction is therefore dissipated by the currents flowing along the bottom of the oceans. Eventually the particles react with oxygen dissolved in seawater. The oxides of iron and manganese are insoluble in water; hence they form metalliferous sediments. The sediments include a certain amount of metallic copper, zinc, cobalt and nickel. Nevertheless, they are an ore too poor for commercial exploitation.

Precisely where is metal-rich sediment deposited? In the wake of Boström's discovery (with Peterson) that metalliferous sediments are ubiquitous on the ocean floor, he undertook to map the worldwide distribution of the sediments' metal content. On the floor of the South Pacific he found some peculiar patterns. At 15 degrees south latitude a long, narrow salient of sediment relatively rich in metal projected westward from the South Pacific ridge axis. At 30 degrees south latitude a broader salient projected eastward. Later John E. Lupton and Harmon Craig of the Scripps Institution mapped a spectacular plume of helium 3 extending westward from the ridge axis across much of the width of the Pacific at 15 degrees south latitude.

These various observations went without an explanation until a recent publication by Joseph L. Reid of Scripps. He had been surveying the density of the water from place to place in the oceans. From the densities he derives the shear forces between ocean layers. Then, having posited that the ocean has "surfaces" where water stands still (such surfaces are boundaries between stable currents flowing in different directions), he can calculate the magnitudes and directions of flow with respect to a given surface. Reid has developed maps of the motion of the water at a depth of 2,000 meters in the South Pacific with respect to a "level of no motion" that is taken to lie at 3,500 meters. Remarkably, the asymmetries found by Boström are exactly coincident with the flow lines inferred by Reid. The plume of helium 3 also stands explained: Helium is an unreactive chemical element; therefore it undergoes no change as it moves with ocean currents. In effect it labels the ocean water from which the hydrothermal effluent rains down on the ocean floor.



BESSHI DEPOSIT, an ore body in the midst of sedimentary rock, is thought to form under the ocean when a ridge axis close to a land mass is buried by erosional debris. Metal-rich sulfides aggregate hundreds of meters above basaltic dikes and sills: vertical and horizontal intrusions of magma that generate hydrothermal activity as they cool after entering the sediment.

Besshi Deposits

If the discoveries at the ridge axes are contributing greatly to the understanding of the processes that form ophiolite ore bodies, they are contributing with equal consequence to the understanding of a type of continental deposit of greater economic importance. It is the Besshi deposit, named after the one found in Japan. In a Besshi deposit a massive sulfide ore body lies in the midst of sediment, usually shale, a rock that originates as a fine-grained clay. The ore body does not lie on top of basalt. In fact, one must often drill in the shale to a depth of several hundred meters under the ore before any basalt is found, and even then the basalt takes the form of dikes and sills: vertical and horizontal intrusions of magma.

Evidently a Besshi deposit forms only when a ridge axis is close to a land mass that acts as a source of large amounts of erosional debris. One example is known in which the process is under way. At the Guaymas Basin in the middle of the Gulf of California the East Pacific Rise is penetrating into a continental plate. River sediment from Mexico has buried the axis in silt to a depth of several hundred meters. Under these circumstances the emplacement of newly minted oceanic crust takes an idiosyncratic course. The magma ascending at the site does not flow onto the ocean bottom. Instead it rises into the silt. The result is dikes and sills. The silt is quite permeable to water; hence hydrothermal activity is widespread. Indeed, the site was first identified (by Lawrence A. Lawver of Scripps) because of the great submarine flux of heat he had measured in the area.

Subsequently Peter F. Lonsdale of Scripps found sulfide mounds, each several tens of meters high and hundreds of meters long, at the center of the Guaymas Basin. The melting of the plastic liners of Lonsdale's core samples suggested a temperature of 100 degrees C. as little as 10 meters under the ocean floor. John Lupton of Scripps found high concentrations of helium 3 in the ocean above the site. Deep drilling penetrated hot basaltic sills.

In January of last year we took part in an expedition that Lonsdale led to the area. In seven dives *Alvin* sampled water venting from three of the largest mounds. The hottest waters had a temperature of 315 degrees C.; their very low concentration of magnesium and sulfate identified them as the end member of the system. In general, however, their composition made them dramatically different from the end members vented at ridge axes in the open ocean. Doubtless their ascent through the Guaymas silt is responsible for the difference. In particular the calcium carbonate shells of plankton buried in the silt are dissolved by the Guaymas end member, which as it leaves the basaltic reaction zone must be quite acidic (like the end members issuing from black smokers at 21 degrees north latitude). The addition of the carbonate, which makes the solution alkaline, induces the precipitation of sulfide minerals. By the time the Guaymas end member enters the ocean it is about four times more alkaline than seawater and is nearly depleted in iron and the other ore-forming elements.

The Smell of Diesel

At Guaymas, then, the crucial difference is really the richness of planktonic life in the overlying ocean. At Guaymas the end member encounters alkalinity under the ocean floor and so the metal content of the solution is quite efficiently trapped. On an open ridge axis, such as the one at 21 degrees north latitude, the end member encounters alkalinity only when it meets seawater. The result is black smoke and the dispersal of much of the end member's metal content.

At Guaymas the richness of plankton seems to have worked a curious modification of the vent-field ecosystem. When Lonsdale collected sediment cores in the Guaymas Basin, he noted that they smelled strongly like diesel fuel. When we returned last year, the water samplers on *Alvin* got clogged repeatedly with globules of wax. These findings are readily explained. The Guaymas silt is rich in planktonic carbon, which is "cracked" into hydrocarbons by the heat of the hydrothermal fields. The Guaymas fields are notable for great mats of bacteria; presumably the bacteria feed on the hydrocarbons. The bacteria appear to compete with clusters of tube-dwelling worms around the vents. Clams, mussels, fish and small crabs are missing. On the other hand, large horned crabs are common; they seem to prey on the worms. In all respects the Guaymas vent fields are different from the open ridge-axis fields as manifestations of hydrothermal activity. They hint at the variety to be expected in future discoveries.

The past two decades have seen an explosion in the earth and planetary sciences. On the earth the theory of plate tectonics has unified the hitherto fragmented and descriptive science of geology. Furthermore, the discovery of hydrothermal activity at the submarine ridge axes demonstrates that the new and the unique are not the exclusive province of spacecraft missions to other planets. It is said that the earth is the water planet. Now it can be said that the chemistry of the water and of ocean sediment bears the prominent imprint of volcanic processes.

A Window on the Sleeping Brain

REM (rapid eye movement) sleep, the phase of sleep when vivid dreams occur, is normally accompanied by paralysis. The paralysis can now be turned off in animals, making it possible to explore the REM phase

by Adrian R. Morrison

Sleep is ordinarily regarded as a condition of complete relaxation and inactivity. When the electrical activity of the brain in a sleeping human being or experimental animal is observed, however, it becomes evident that sleep is a complex and by no means inactive state. Indeed, the electrical activity of the brain in the phase of sleep when vivid dreams occur—REM (rapid eye movement) sleep—resembles that of wakefulness more closely than it does that of the other phases of sleep. In REM sleep transitory voltage changes indicate that the brain is in a state of arousal, in spite of the fact that sensory contact between the animal and its environment is much reduced. What prevents the neural arousal of the brain in REM sleep from being translated into vigorous physical movement is the absence of activity in the animal's muscles, which results in a paralysis that lasts until an episode of REM sleep ends.

What would happen if the paralysis that accompanies REM sleep were eliminated and the aroused brain were enabled to activate the muscles? Work in my laboratory at the University of Pennsylvania and in other laboratories is beginning to answer the question. The most primitive part of the brain is the brain stem, which lies between the spinal cord and the rest of the brain. In the cat small, precisely controlled lesions in the brain stem result in animals that show the electrical characteristics of REM sleep but move vigorously. The lesions are made in the region of the brain stem called the pons, and the exact position of the lesion within the pons strongly affects which muscles are freed from paralysis.

Therefore the lack of muscle tone in sleep appears to be directly attributable to the pons. Other experimental results suggest that neurons (nerve cells) in the pons also affect locomotion indirectly. In REM sleep one region of the pons appears to interfere with the action of a "locomotion center" that extends from the pons to a point farther down the brain stem. This interference prevents the locomotion center from activating the nerve networks in the spinal cord that are responsible for the reciprocal (alternating) motion of the limbs. Thus the activity of the pons in REM sleep leads to the inhibition of movement in two ways. My own work suggests that the link between heightened brain activity and muscular paralysis can also operate in wakefulness and may be the cause of certain sleep disorders.

The transition from alert wakefulness to sleep entails many changes in an animal's physiology and posture; the alterations collectively result in a profound change of state. Much of my work has been done with cats, and the changes described here apply specifically to the cat, but many of them are shared by other mammalian species. When a cat falls asleep, it assumes a familiar curled-up and relaxed pose. Its eyes close, and the nictitating membrane (the "third eyelid") covers part of the eye under the outer eyelids. There is a reduction in the number of muscle fibers that are contracting and therefore a reduction in the degree of tone in the muscles. In the initial phase of sleep the movements of the eye that can be detected in the waking animal disappear. There is a gradual lowering of brain temperature, generally by a fraction of a degree Celsius.

Underlying such changes are the changes in the neural activity of the brain that occur when sleep begins. The most general information about the electrical activity of the brain is provided by the electroencephalogram (EEG). The EEG is a record of continuous changes in electric potential measured between electrodes affixed to the skull. The rapid changes in potential represent the sum of changes in voltage in the outer membrane of large numbers of brain neurons.

The EEG in an awake and alert cat shows waves with a relatively low amplitude and a relatively high frequency. As the cat curls up and goes to sleep there is a notable change. The low-amplitude waves of wakefulness are replaced gradually by high-amplitude, low-frequency waves. Because of the presence of the low-frequency pattern in the EEG, the light initial phase of sleep is often referred to as slow-wave sleep. As the animal enters slow-wave sleep there are also changes in electrical activity in specific areas of the upper and lower brain. Such changes are recorded by means of electrodes inserted into the appropriate brain structure.

For many years the slow-wave pattern was thought to prevail throughout sleep. In 1953, however, Eugene Aserinsky and Nathaniel Kleitman of the University of Chicago noted in human beings intervals of sleep when the EEG record reverted to the high-frequency, low-amplitude pattern of wakefulness. After the EEG pattern had changed, the sleepers' eyes were periodically observed to move rapidly in various directions; it is for this reason that the intervals have been designated rapid-eye-movement sleep. Periods of REM sleep are now known to recur regularly during sleep, alternating with longer periods of slow-wave sleep. The timing and the duration of the REM periods vary with the species. REM sleep comes about once every 90 minutes in human beings and about once every 25 minutes in cats. Each episode lasts for several minutes.

By waking sleepers immediately after an episode of rapid eye movement had ended Aserinsky and Kleitman were able to show that REM sleep is associated with vivid and intense dreams. The identification of the time in sleep when the sleeper dreams was the source of much of the initial excitement over the discovery of REM sleep. Later work has shown, however, that the physiological aspects of REM sleep are equally intriguing.

In addition to the essentially identical EEG patterns in REM sleep and in the waking state the rates of activity of neurons in most of the subunits of the brain are quite similar in the two states. One of the most striking similarities is found in the hippocampus, which is phylogenetically the oldest part of the cerebral cortex. Throughout much of both REM sleep and wakefulness there is a regular pattern of waves in the hippocampus at a rate of about seven per second. This pattern, the "theta rhythm," is quite different from the activity of the hippocampus in slow-wave sleep, which is less regular and shows spikelike waves. An additional similarity is in brain temperature: after decreasing part of a degree in slow-wave sleep the temperature of the brain increases to roughly the waking level during an episode of REM sleep.

Hence many of the characteristics of REM sleep are parallel to those of wakefulness and are quite different from the characteristics of slow-wave sleep. Remarkably, all the characteristics mentioned above are signs of activation of the reticular formation, or reticular activating system, in the core of the brain stem. In 1949 Giuseppe Moruzzi and Horace W. Magoun, working together at Northwestern University, demonstrated that the reticular formation is responsible for arousal. Thus in REM sleep the brain is highly aroused. Because of the juxtaposition of rest and arousal, REM sleep is sometimes called paradoxical sleep.

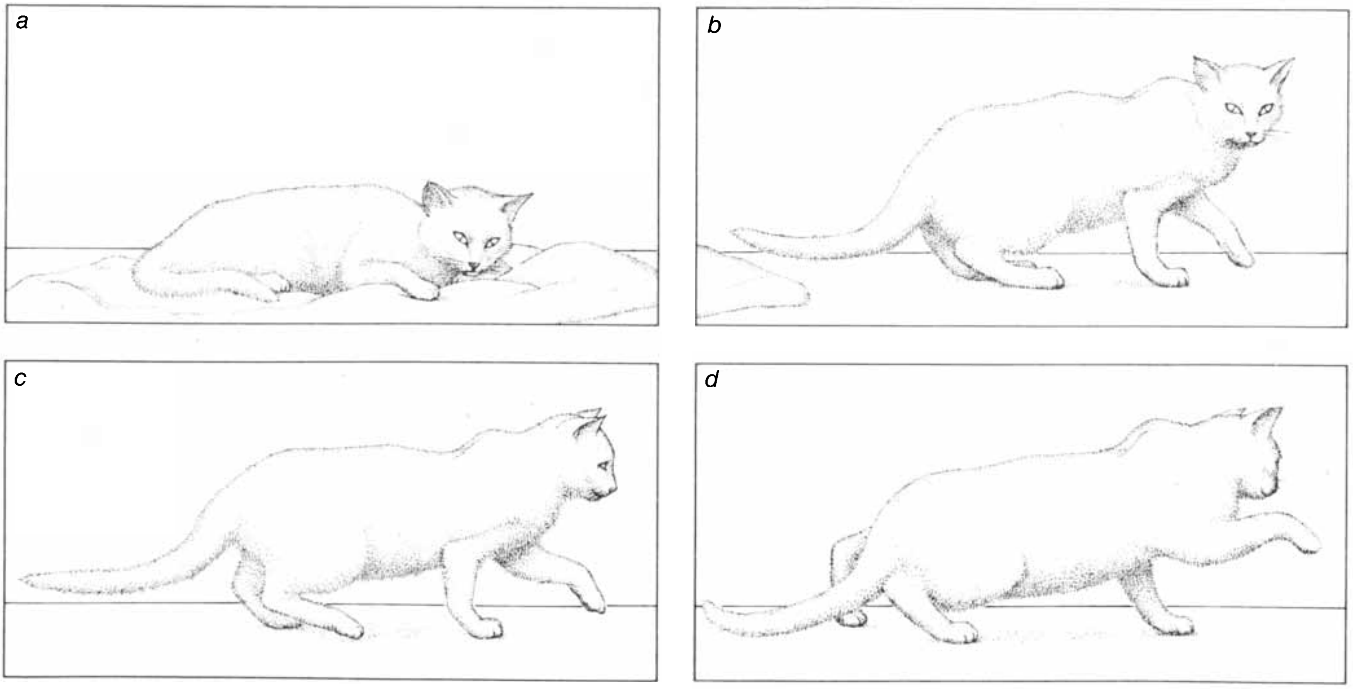
The close connection between REM sleep and alert wakefulness was emphasized recently by an unexpected finding in my laboratory. One of my students, Robert Bowker, had been working on electrical waves in REM sleep that appear "spontaneously" (meaning without apparent external stimulation). These brief, high-amplitude waves are called PGO spikes in reference to the brain structures in which they have been most studied: the pons (where they are thought to originate) and two parts of the visual system, the lateral geniculate body and the visual cortex (the occipital region of the cerebral cortex).

It had been thought that PGO spikes were limited almost exclusively to REM sleep and occurred only rarely in slow-wave sleep. One day in the laboratory, however, Bowker accidentally tapped on the recording cage while a record was being made from the brain of a cat in slow-wave sleep. Almost immediately a PGO spike appeared in the record. Further work showed that PGO spikes could be readily elicited in either REM sleep or slow-wave sleep by sounds and by tactile stimuli. Hence PGO spikes, which were thought to be spontaneous electrical events occurring only in REM sleep, were seen to be multisensory alerting responses that can be evoked in several brain states.

The conclusion that PGO spikes are alerting responses led Bowker to reexamine the waves called eye-movement potentials (EMP) that appear in the waking state. Such waves in the electrical record are identical in appearance with PGO spikes and are observed in the same brain structures. The presence of eye-movement potentials, however, was thought to depend on the level of illumination in the cat's environment. Such potentials had not been observed in awake cats in the dark. The dependence of eye-movement potentials on illumination was thought to constitute a significant difference from PGO spikes.

Reasoning that a cat in a dark recording chamber might be somewhat bored, Bowker aroused the animals' interest by sending the odor of tuna fish wafting through the cage. Soon the cat's record showed eye-movement potentials that were identical with PGO spikes. Short bursts of a loud tone yielded the same result. The result suggests that both PGO spikes and eye-movement potentials are particular forms of a general alerting response. The response can be elicited by stimulation in the waking state, in slow-wave sleep or in REM sleep. Activation of the same neurons can apparently also occur spontaneously in REM sleep. Thus in an episode of REM sleep the brain is aroused to the extent that it is in a state resembling alert wakefulness even though sensory contact with the environment is greatly reduced.

What prevents the animal in REM sleep, whose central nervous system is aroused and alert but for the most part cut off from sensory contact with the world, from injuring itself? The answer is that in each period of REM sleep the action of the motor neurons of the spinal cord that cause the skeletal (voluntary) muscles to contract is inhibited.



REM SLEEP WITHOUT PARALYSIS is shown in drawings based on motion pictures of a cat made in the author's laboratory at the University of Pennsylvania. The paralysis that normally accompanies REM sleep is eliminated by making a small lesion in the brain stem of the cat; the degree to which the paralysis is eliminated varies with the lesion. The cat in the filmed episode raised its head and righted itself during REM sleep (*a*). It managed to stand (*b*) and began to walk (*c*). It then began to paw at the air in a corner of the laboratory (*d*). In such episodes the cat shows all the signs of normal REM sleep, including the extension of the nictitating membrane (the "third eyelid") to cover much of the eye. That not all the muscle tone is released is suggested by the cat's slightly swaybacked appearance.

As a result the muscles are atonic, or without tone, and the animal is paralyzed. (It should be noted that atonia and paralysis are not identical. Paralysis can be the result of several conditions other than atonia. If the muscles are atonic, however, the animal will necessarily be paralyzed.)

The degree of activity in the skeletal muscles is often recorded by means of an electrode inserted into the neck. The neck muscles are among the "antigravity" muscles that are responsible for the maintenance of upright posture. As we have seen, when the cat enters light sleep, the tone in its muscles decreases because there is a decrease in the number of active fibers in each muscle. When the animal enters a period of REM sleep, the record of muscle tone goes flat, indicating a complete lack of tone. There are occasional bursts of muscle activity, however, that result in twitches in various parts of the body. In my laboratory we have been working with cats in which the atonia of REM sleep has been reversed by means of lesions in the pons made by destroying a small volume of tissue with a heated wire. The wire is introduced into the brain at predetermined coordinates.

After such a lesion is made, remarkable activity is sometimes observed. The activity follows a period of slow-wave sleep, at the time when the cat would ordinarily enter REM sleep with atonia. The cats raise their head, right their body and exhibit alternating movements of their limbs. They attempt to stand, and some succeed; others are actually able to walk. Cats with certain lesions display other behavior usually seen in wakefulness. They make motions typical of orienting toward prey, searching for prey and attacking. These movements are rarely directed at anything in the environment.

As I have noted, the kind of muscular activity the cat displays in these extraordinary episodes depends on the size and position of the brain lesion. When sections from the brain of a cat that could support itself on forelimbs only were observed with the light microscope, small, symmetrically placed lesions were noted in the dorsal, or upper, part of the pons. A cat that could support itself on all four limbs and walk was observed to have larger lesions in a more ventral, or lower, position. A cat that showed aggressive behavior, striking repeatedly at the floor in front of it, had lesions that extended forward into the midbrain. The significance of these positions will become clearer when we examine the neural pathways implicated in the inhibition of movement during REM sleep.

Several items of information suggest that the unusual movements we observe do indeed constitute REM sleep without atonia. The phenomenon was initially recognized by Michel Jouvet and François Delorme at the University of Lyons. As in ordinary REM sleep, the EEG of the cats that are active in sleep resembles that of wakefulness. The nictitating membrane partially covers the eye and the pupils close to slits. Recordings made from the hippocampus show the theta rhythm. Moreover, my student Joan Hendricks confined the sleeping cats in a padded harness, and at times when the cats with lesions would have been engaging in complex movements if they had been free they showed the rapid movements of eyes, whiskers and digits normally seen in REM sleep. Hendricks found that the temperature of the brain rose during these episodes, as it does in ordinary REM sleep.

Hendricks also showed that one particularly intriguing quality of REM sleep is present in REM sleep without atonia. Pier Luigi Parmiggiani and his colleagues at the University of Bologna demonstrated that the capacity of the cat to regulate its body temperature is suspended in REM sleep. In REM sleep cats appear to be "cold-blooded," like fishes, amphibians and reptiles. Hendricks showed that in REM sleep without atonia cats do not shiver or fluff up their coat in response to cold, although the same animals do show such responses when they are awake. Thus the lesions in the pons do not restore the capacity to respond to cold.

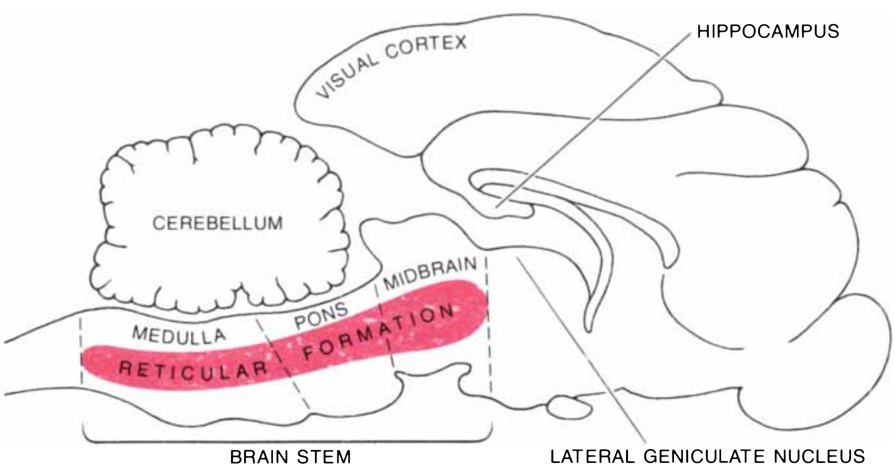
Much of my current work is aimed at understanding how the small lesions in the pons cause the atonia of REM sleep to be reversed and lead to elaborate movements that appear to be much like those of a waking animal. We have concentrated on the pons because in 1962, early in his work on sleep, Jouvet demonstrated that cats deprived of the entire brain in front of the pons still showed periods of atonia and rapid eye movements that were identical with those of REM sleep. In the course of my work I have concluded that there are at least two separate neural systems operating in REM sleep without atonia. One system is responsible for the release of muscle tone, the other for the liberation of "motor drive," a generalized impulse toward locomotion.

The first system appears to be the simpler of the two. It is known that an inhibitory center in the medulla is operative in REM sleep. (The action of neurons can be either excitatory or inhibitory. Inhibitory areas reduce the excitability of the target neurons.) In REM sleep the inhibitory center in the medulla interferes with the spinal neurons that activate the skeletal muscles. This mechanism is thought to be responsible for the lack of tone in the skeletal muscles.

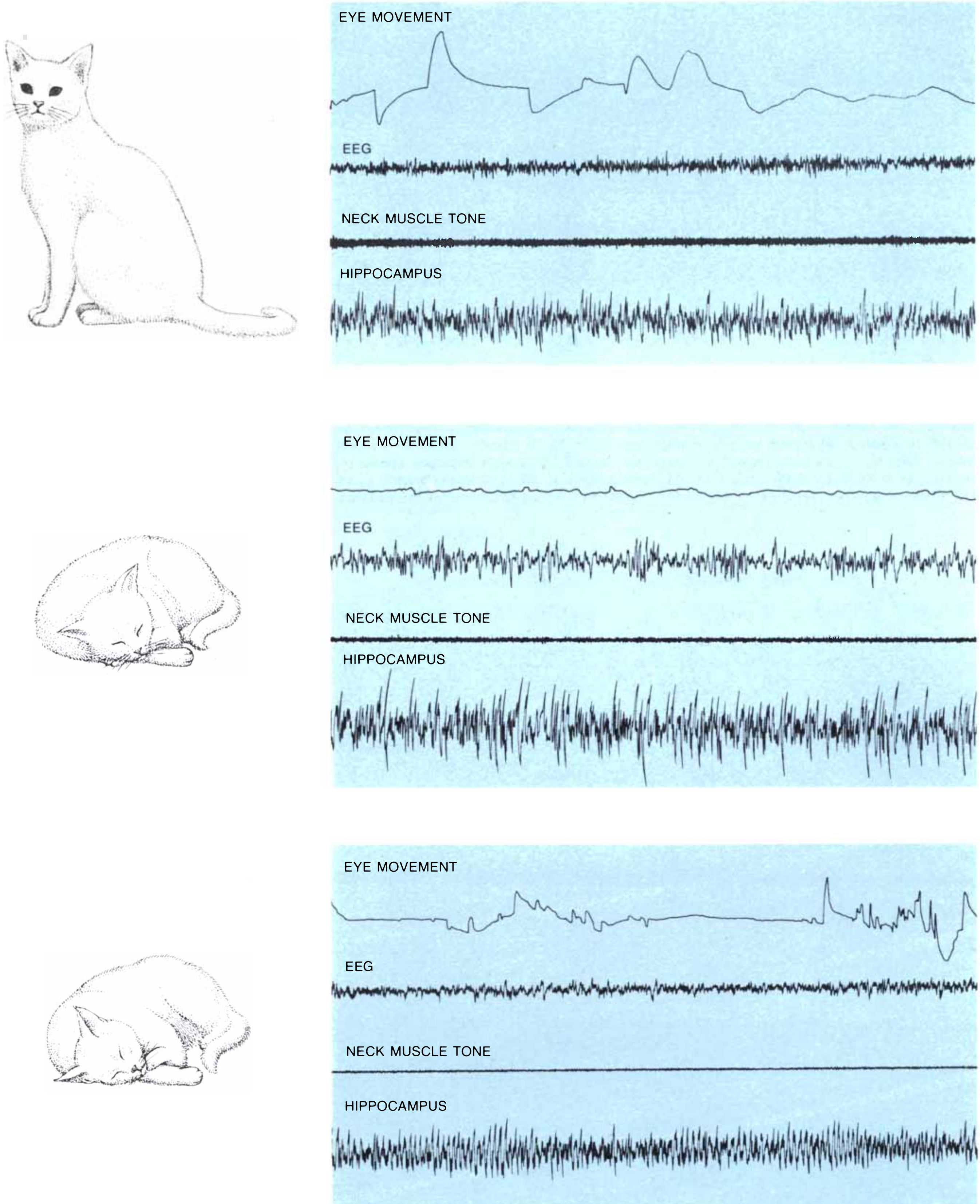
If the inhibitory center in the medulla were in turn under the excitatory control of a network of nerve cells in the pons, damage to the pons could break the excitatory connection. The action of the inhibitory center in the medulla would be interrupted and the skeletal muscles would retain their tone in REM sleep. The extent of the release of tone, and hence the muscles the cats could employ, would depend on the position of the lesions.

This hypothesis has received support from results obtained by Jouvet and his colleagues at Lyons. They have shown that a tract of nerve fibers originates in an area just dorsal to the area in the pons where we made lesions. The tract leads directly to the inhibitory area in the medulla. We have made lesions in the origin and in the course of the tract. Such damage yields cats that have some neck muscle tone in REM sleep but can lift their head only a little. This suggests that other neurons, probably in the reticular formation, must be damaged to release muscle tone completely.

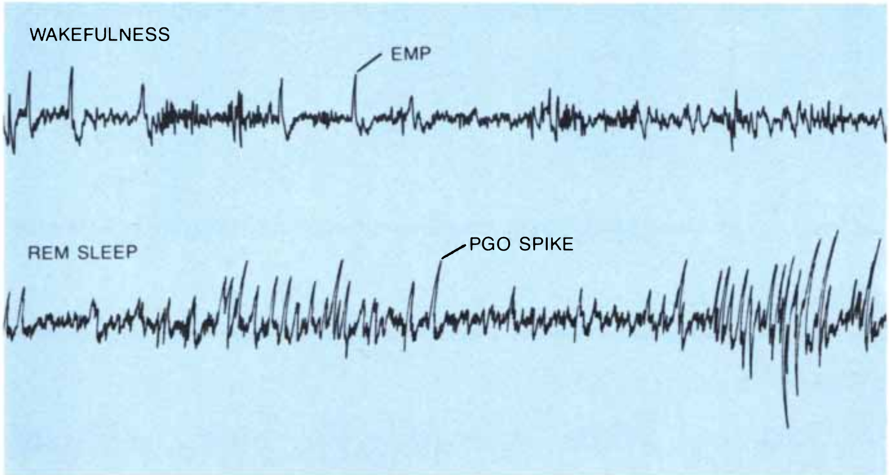
Therefore some experimental evi-



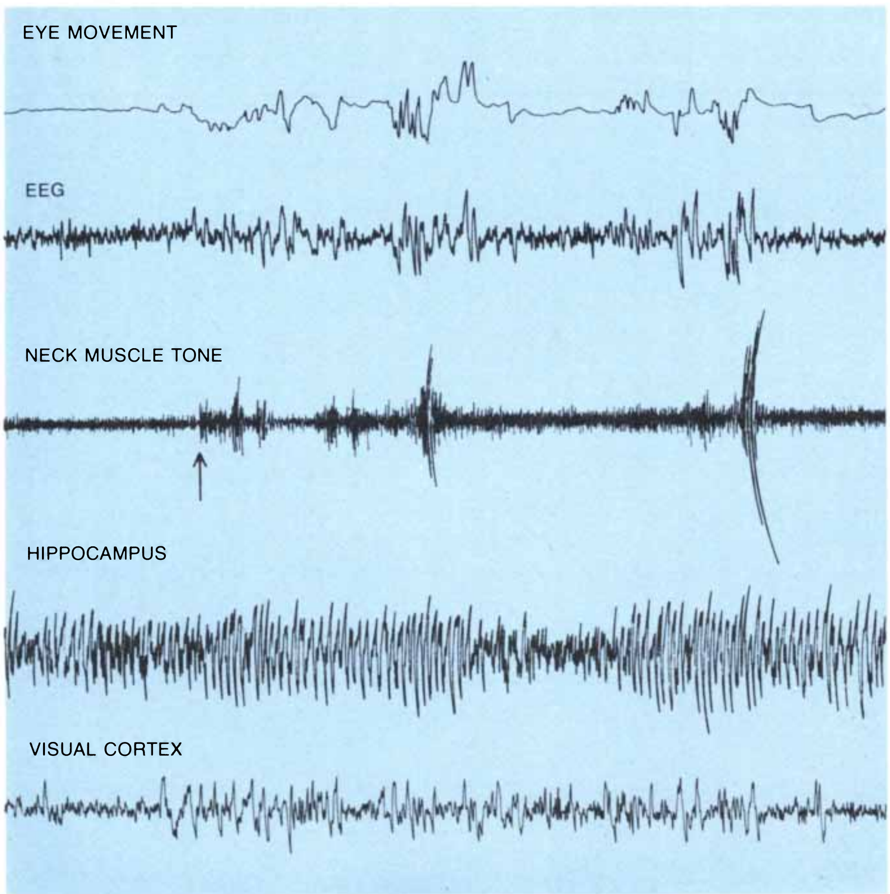
BRAIN STRUCTURES that play a role in sleep are indicated in a schematic section of the brain of a cat. The front of the brain is to the right. The lesions that eliminate paralysis in REM sleep are in the pons, a structure in the brain stem between the medulla and the midbrain.



THREE STATES (wakefulness, slow-wave sleep and REM sleep) are detectable in the cat by changes in posture, muscle tone and brain activity. In wakefulness (*top*) the animal's eyes and head move as the animal responds to visual and auditory stimuli. The electroencephalogram (EEG), which records changes in voltage between electrodes put on the skull, shows a pattern of low-amplitude, high-frequency waves. There is substantial tone in the skeletal (voluntary) muscles, which is usually measured in the neck. The hippocampus, a primitive area of the cerebral cortex, shows a regular pattern called the theta rhythm when the cat is attentive to an object. As the animal enters slow-wave sleep (*middle*) it curls up. Eye movements cease. The EEG shows a low-frequency, high-amplitude pattern. Muscle tone decreases and the hippocampus shows an irregular rhythm. When the animal enters a period of REM sleep (*bottom*), which in the cat comes about once every 25 minutes during sleep, the curled-up posture relaxes slightly. Rapid eye movements, a high-frequency EEG pattern and the theta rhythm reappear. Skeletal muscle tone, however, disappears completely, a condition that is referred to as atonia.



BRAIN AROUSAL IN REM SLEEP was demonstrated recently in the author's laboratory. In the waking state waves called eye-movement potentials (EMP) appear in the visual area of the cerebral cortex (*top*). Such potentials had been thought to depend solely on the level of illumination. The author's student Robert Bowker showed that eye-movement potentials can be elicited by a variety of stimuli even in absolute darkness: they are multisensory alerting responses. Such waves are identical with the waves known as PGO spikes that appear spontaneously in the same areas of the brain in REM sleep (*bottom*). In REM sleep the brain is alert. The alertness, however, results from the brain's activity rather than from sensory information.



EPISODES OF MOVEMENT DURING SLEEP in the cats with lesions take place when all signs except muscle tone are similar to those of REM sleep. The similarities have led the author to conclude that the episodes of movement represent REM sleep without the usual atonia. Recordings from normal REM sleep include rapid eye movement, low-amplitude EEG pattern, theta rhythm and PGO spikes. All are present in recordings such as the one shown, made during the episodes with movement that follow damage to the pons. The notable difference is that in REM sleep without atonia the cat's skeletal muscles have regained much tone, enabling the animal to move. The arrow indicates the start of an episode of movement during REM sleep.

dence bears out the notion that a fairly direct connection between the pons, the inhibitory center in the medulla and the skeletal muscles underlies the atonia of REM sleep. Since REM sleep in human beings is known to be associated with intense dreams, it would be tempting to conclude that by breaking this connection we are enabled to witness the animal acting out its dreams. Apart from the difficulties inherent in attributing complex mental states to other species, however, there are good reasons to think this is not a full account of what happens in the episodes of REM sleep without atonia.

The functions of the brain are carried out by closely interconnected anatomical structures. It would be unreasonable to expect that the only result of damaging a central area of the brain stem would be to destroy structures inhibiting motor neurons. Indeed, we have evidence that systems other than the inhibitory one in the medulla are affected by the lesions in the pons.

The most important evidence concerns the overall level of locomotor activity when the cats with lesions are awake. The cats show no abnormal increase in muscle tone in slow-wave sleep or in wakefulness; the only effect on muscle tone is in REM sleep. Nevertheless, the cats appear to be more active in general than they were before the lesions were made. When they are loose in the laboratory, they make something of a nuisance of themselves, running here and there to investigate inconsequential things.

In order to confirm the impression of unusual locomotor activity we conducted open-field tests. Several cats were tested before and after the lesions were made. The cats were put in a room with a floor that had been marked off in squares. The number of squares each cat entered in a 30-minute period served as a measure of locomotor activity. All the cats that displayed REM sleep without atonia also demonstrated increased activity after the lesions. The increases ranged from 30 to 261 percent and were all significant in a statistical sense. This finding led me to conclude that the lesions in the pons affect a source of generalized motor drive. It is probably an anatomical system different from the one that affects muscle tone, since when the animals were awake, their muscle tone was not affected by the lesions.

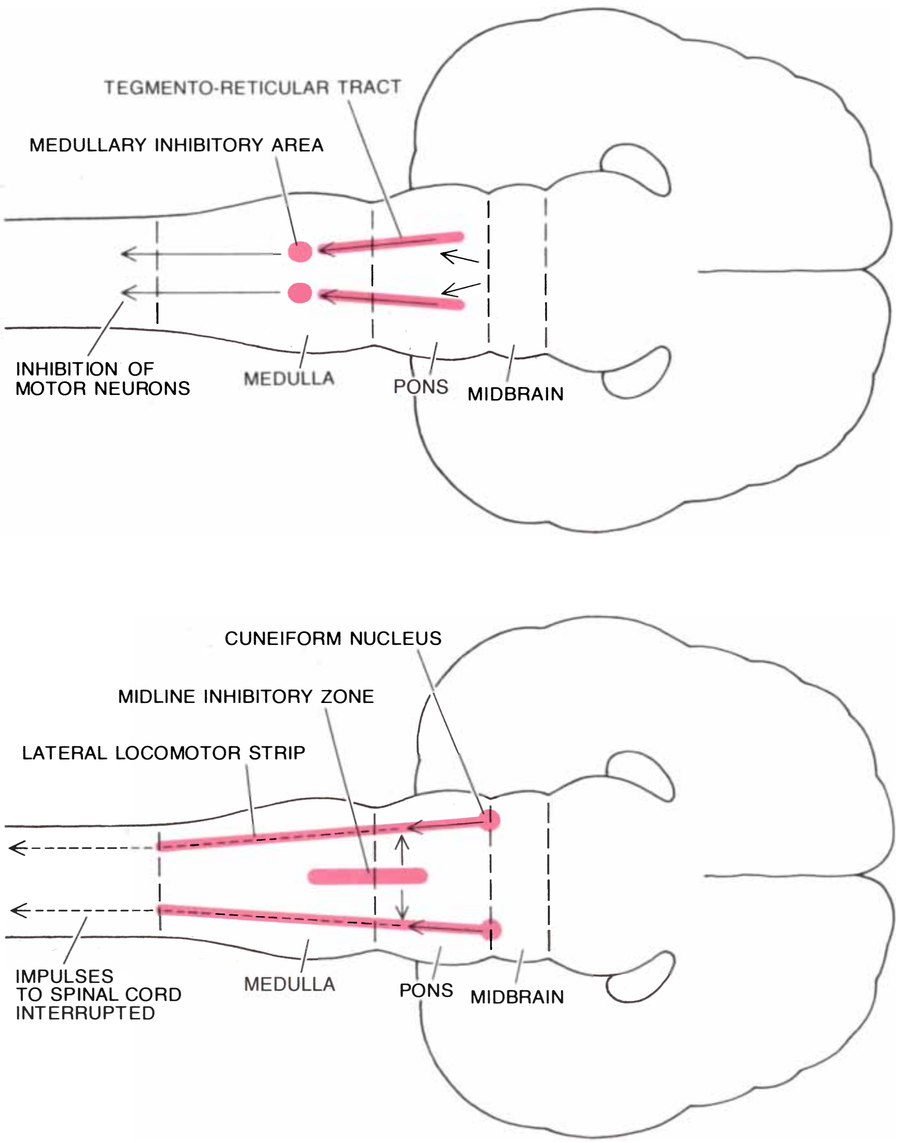
Our work is beginning to yield a picture of how the second system of inhibition in sleep could work. My hypothesis of how the system operates borrows heavily from work done on the neural control of locomotion. It has been shown that the region of the brain stem we are interested in can regulate locomotion without any contribution from higher brain centers. Such regulation involves three structures. A "locomotion generator" in the spine includes neurons that control and coordinate the reciprocal motion of the limbs in walking and running. The generator is under the excitatory control of a second structure: a "locomotion center" in the brain stem. If the locomotion center is stimulated, it in turn stimulates the generator of reciprocal motion. The third structure is a control region in the pons, which is connected to the locomotion center by an inhibitory link. When the control region is activated, it suppresses the activity in the locomotion center. Thus in REM sleep the pons could indirectly suppress motor drive as well as muscle tone. Conversely, damage to the pons could release the muscles utilized in locomotion.

The existence of the brain-stem locomotion center and the spinal movement generator has been postulated by many workers, but their anatomical organization is not yet understood. The first evidence of the existence of the locomotion center was obtained by M. L. Shik, F. V. Severin and G. N. Orlovskii of the Moscow State University; their results have since been confirmed by other workers. Shik and his colleagues have shown that the application of an electric current to a neural system originating in the cuneiform nucleus, which is in the caudal region of the midbrain, can cause cats that have been deprived of the cerebrum to walk on a moving treadmill and even to trot or gallop. I suspect it is this center that the lesions in the pons release from inhibition in REM sleep.

Recent work by Shigemi Mori and his co-workers at the Ashikawa Medical College in Japan supports the existence of the locomotion center and suggests how the center may be connected to the pons. The Japanese workers have found that the tracts capable of inducing locomotion in cats deprived of the cerebrum extend from the midbrain down the outer part of the brain stem, in the form of two thin strips passing down the sides of the pons and the medulla. Moreover, they have shown that there are areas in the central part of the brain stem capable of facilitating or inhibiting the operation of the locomotion center.

Whether the locomotion center is excited or inhibited depends on the precise area that is stimulated. Electrical stimulation of the more ventral facilitatory zone is capable of inducing reciprocal movement of the animal's hind legs without any direct stimulation of the locomotion center itself. Dorsal stimulation reduces muscle tone and inhibits locomotion. The facilitatory and inhibitory zones discovered by Mori and his colleagues are near the midline of the brain stem and are close to the regions in the pons where our lesions are made. Thus the connection between the pons and the putative locomotion center is becoming clearer.

Such findings have implications for interpreting the effects of the lesions in the pons. For example, in the cat mentioned above that was able to stand and walk, the lesion was capable of cutting off the full effect of the midline inhibitory zone on the locomotion center. In the cat that could raise itself on its forelimbs only, the lesion was not capable of interrupting the full midline inhibitory effect. My associate Graziella Mann and I have recently found that additional damage to the inhibitory zone in a cat capable of raising itself on forelimbs only can enable the cat to walk on all fours in REM sleep without atonia. Nevertheless, the anatomical connections in the pons and between the locomotion center and the spinal generator remain speculative. Further work in this area is much needed. In the cats that show aggressive behavior there are additional complications. Such animals always have damage in the neural path-



SEPARATE NETWORKS of neurons (nerve cells) appear to be responsible for inhibiting muscle tone and locomotor drive in REM sleep. The action of neurons on their targets can be either excitatory or inhibitory. An area in the medulla that is known to inhibit motor neurons plays an important role in reducing muscle tone (*upper panel*). The tegmento-reticular tract connects the pons to the inhibitory center. In REM sleep the pons is activated, exciting the medullary center by this pathway and others. The medullary center inhibits the motor neurons and gives rise to atonia. A lateral locomotor strip plays an important part in the reduction of motor drive (*lower panel*). The strip runs down the outside of the brain stem. It is connected to structures in the spinal cord and can induce limb motion in cats from which the cerebrum in front of the midbrain has been removed. In the core of the brain stem are neurons that can excite or inhibit the locomotor strip. They are arranged in facilitatory and inhibitory zones, lying in the same vertical plane near the midline of the brain. In REM sleep the pons stimulates the inhibitory zone, turning off the locomotor strip and shutting down motor drive. Thus damage to the pons can eliminate inhibitory control in both systems, freeing muscle tone and motor drive.

ways descending from the cerebrum that are known to be involved in the control of aggression in wakefulness.

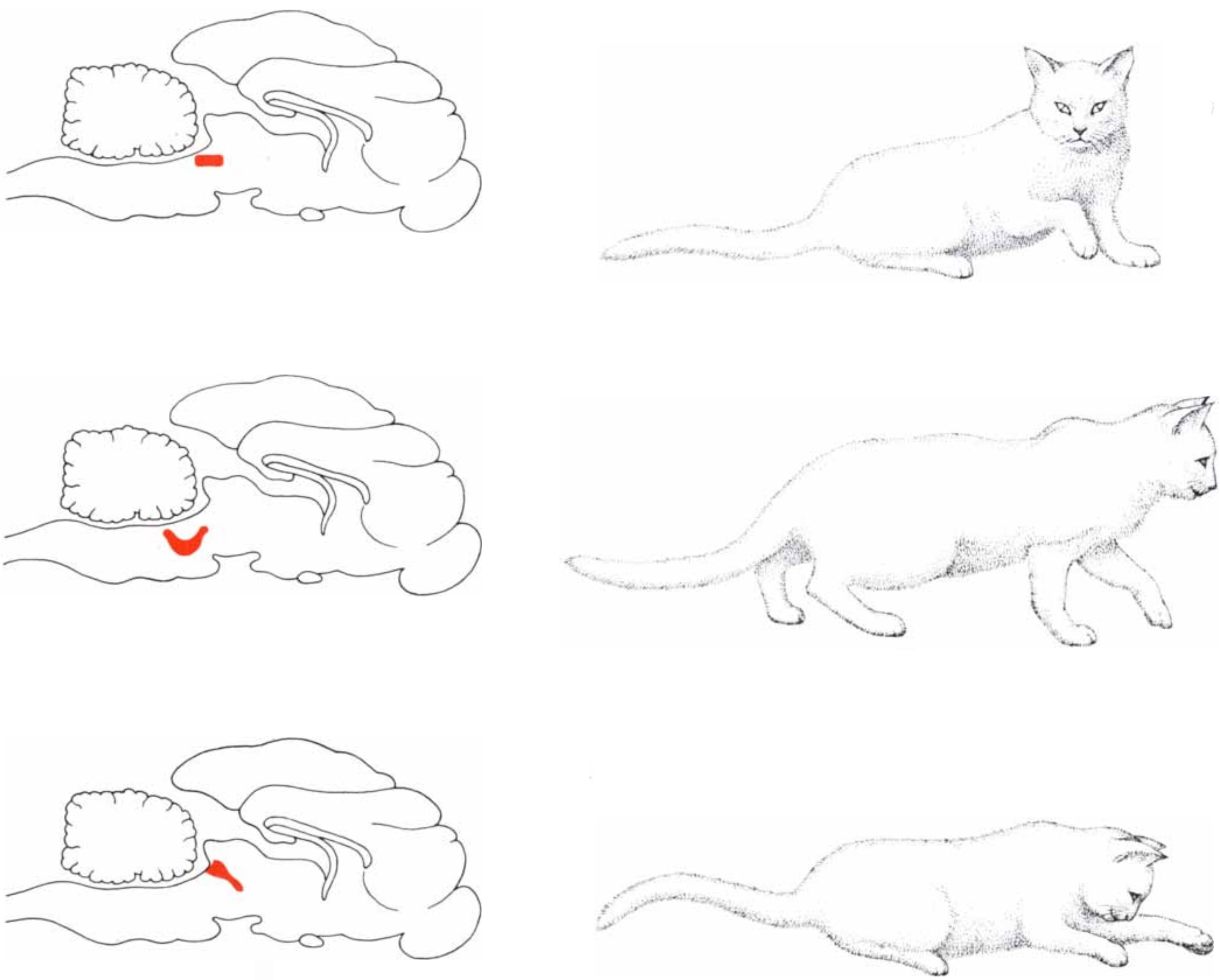
It should not be assumed, however, that even when the specific anatomical connections have been worked out, the control of the muscles in REM sleep will be completely understood. Even subtler mechanisms than the ones visualized here could be operating. For example, René Drucker-Colin of the University of Mexico and his students Gloria Arankowsky and Raul Aguilar and I have recently shown that the high-frequency neuron discharges occurring in many regions of the brain in REM sleep are necessary for the release of activity in REM sleep without atonia.

The administration of chloramphenicol, a common antibiotic, at a dose comparable to that employed in treating bacterial infection causes atonia to reappear in cats with lesions in the pons that would otherwise show episodes of movement in REM sleep. The cats appear to be normal when they are awake. Drucker-Colin had previously demonstrated that chloramphenicol reduces the rate of firing of the neurons in the reticular formation. The mechanism by which the neural activity is reduced is not known. Chloramphenicol, however, is known to be an inhibitor of protein synthesis, and this action could be responsible for the reduction in the firing rate of the neurons. Whatever the mechanism of the drug is, after it is administered the lesioned cats appear to have too little locomotor drive to generate the elaborate activity they would otherwise show in REM sleep.

The work that has been done so far on REM sleep without atonia indicates there is a link between the heightened arousal of REM sleep and the reduction of coordinated motor activity. Obviously an organism is well served by a mechanism that makes it impossible to move when the brain is very active but unresponsive to external stimulation. Teleological arguments need not be invoked, however, to explain the existence of a connection between brain arousal and motor inhibition. Such a connection could well exist even in the waking state.

Faced with a novel or unexpected stimulus, a person or an animal normally hesitates to some degree before acting. Most people have experienced a moment of hesitation or even a slack feeling in the knees when they have seen a fast-moving automobile bearing down on them. Such waking motor inhibition is generally short-lived. Nevertheless, it suggests that even in wakefulness there is a connection between momentarily heightened alertness and reduced motor activity.

The investigation of this link could have at least one important clinical consequence. The disorder called narcolep-



POSITION OF LESIONS in the brain of the cat strongly affects the kind of movement the animal shows in REM sleep without atonia. A cat with a small lesion high in the pons could support itself on its forelegs only (*top*). Such a lesion can interrupt only part of the effect of the midline inhibitory zone on the locomotor strip shown in the preceding illustration. A cat with a larger lesion deeper in the pons stood on all fours and walked (*middle*). Such a lesion can cut off the full effect of the inhibitory zone. A cat with a lesion extending into the midbrain displayed aggressive behavior (*bottom*). Such a lesion affects neural pathways from the cerebrum that regulate aggressive behavior.

sy is characterized by sudden and unpredictable lapses from wakefulness directly into REM sleep or from wakefulness into paralysis without loss of consciousness. Intriguingly, strong stimulation (such as that accompanying anger, laughter, surprise or sexual intercourse) are the commonest causes of narcoleptic seizure. It is possible that in those who suffer from narcolepsy there is an abnormally low threshold for the link between arousal and motor inhibition; common arousing stimuli could thus lead to atonia or to REM sleep in its entirety at inappropriate moments.

In addition to its possible clinical consequences the investigation of REM sleep without atonia will probably yield further information about the nature of REM sleep itself. It has already helped to clarify one central problem. As we have seen, in most regions of the brain the neurons follow the same pattern in the transition from being awake to being asleep. There is a decrease in the rate of activity from wakefulness to slow-wave sleep and then an increase to the rate of the waking state as an episode of REM sleep begins. Among the neurons that show a different pattern are those of the area of the pons known as the dorsal raphe nucleus. These cells utilize the substance serotonin as a transmitter to alter the activity of the neurons to which they are connected. The rate at which the cells fire decreases from about two impulses per second in wakefulness to almost zero in REM sleep after passing through an intermediate level in slow-wave sleep.

It had been suggested that the inactivity of the neurons in the dorsal raphe nucleus is a fundamental characteristic of REM sleep. Indeed, it had been hypothesized that the decrease in activity causes REM sleep to begin. To test these propositions I collaborated with Barry L. Jacobs and Michael Trulson of Princeton University. Jacobs and Trulson had already thoroughly studied the activity of the dorsal raphe neurons in the waking state and in normal sleep. When we recorded the activity of these cells in REM sleep without atonia, however, the results were surprising. In the cats with lesions in the pons the raphe neurons, instead of falling silent, increased their activity again after slow-wave sleep. The increase was to a rate of about one impulse per second and so did not bring the neurons to the rate in wakefulness. It did, however, yield a rate about six times that of normal REM sleep.

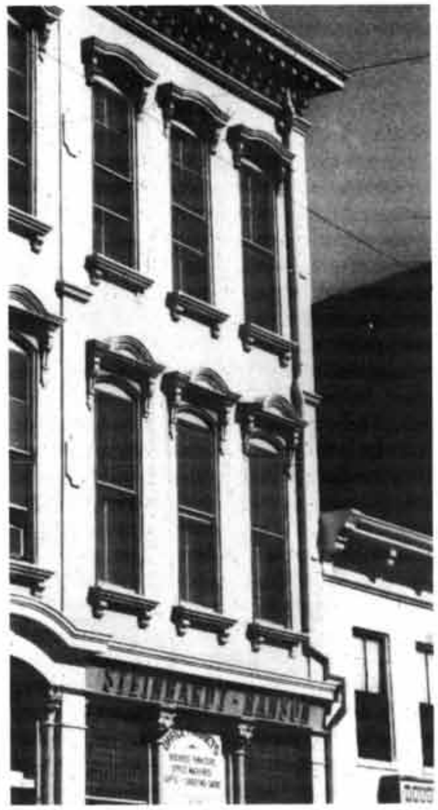
There are at least two plausible explanations for this unexpected finding. The first is that the unusual muscular activity in REM sleep without atonia somehow causes information to be relayed back to the pons to excite the neurons in the dorsal raphe nucleus. An alternative explanation is that the lesions in the pons affect a more central motor mechanism that normally inhibits the raphe neurons in REM sleep. The first hypothesis depends on the relaying of impulses from the periphery of the nervous system to its center whereas the second depends only on events in the brain.

To test the alternative explanations Jacobs and his students employed two drugs. The first is succinylcholine, a drug that is chemically closely related to curare. Succinylcholine acts at the junction between a motor neuron and the muscle cell that the neuron activates. By interrupting the connection the drug induces a temporary paralysis. When succinylcholine was injected into awake normal cats, the raphe neurons remained as active as they had been before the injection. Therefore the raphe neurons could be active in wakefulness with paralysis and in REM sleep without atonia and hence without paralysis. Since the raphe neurons were active in the awake, paralyzed animals, we concluded that the neurons of the raphe nucleus were not being stimulated in REM sleep without atonia by information relayed from the periphery as the result of vigorous muscular activity.

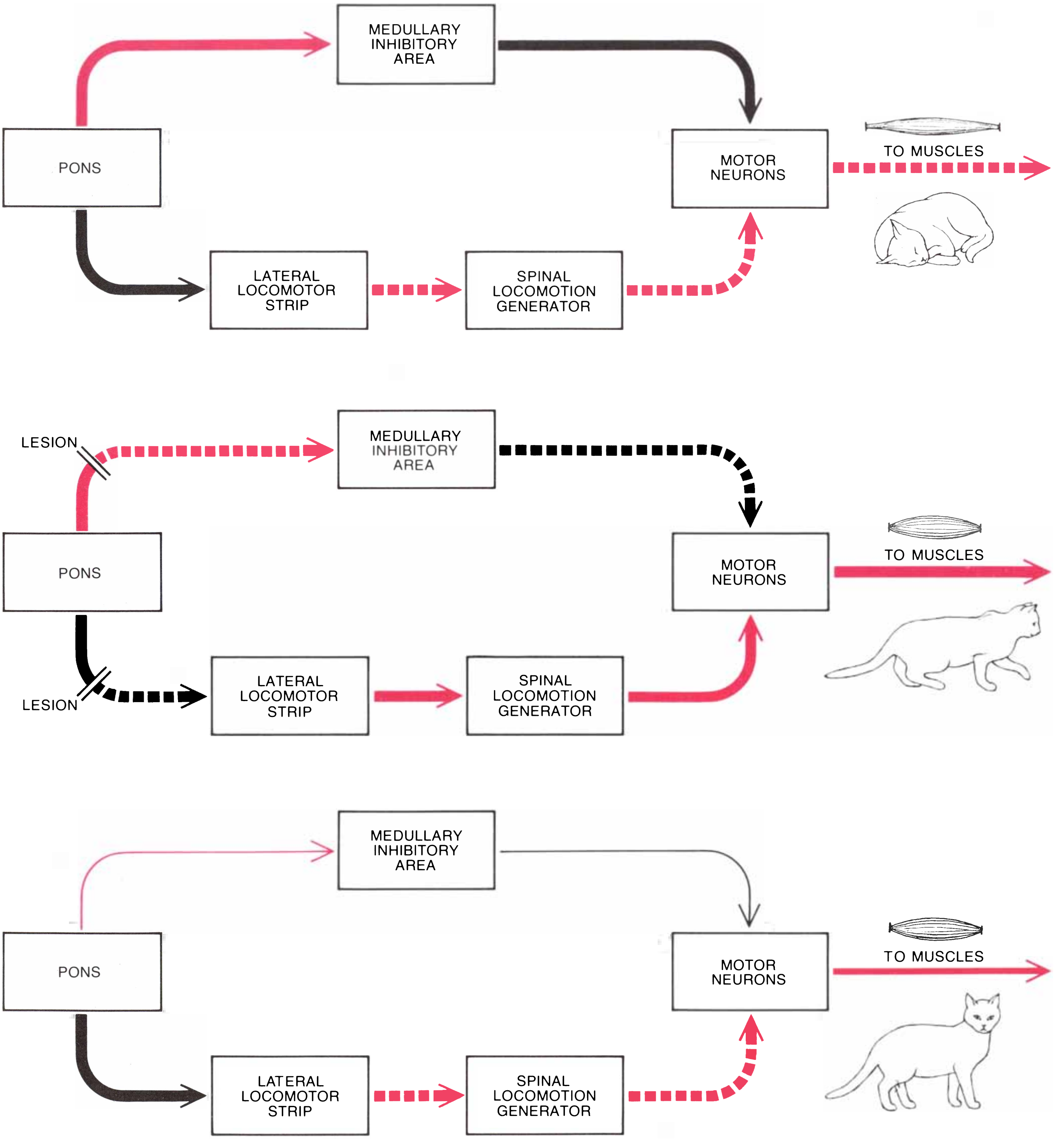
The second drug, carbachol, causes paralysis by directly affecting the mechanism described above that is responsible for atonia. The injection of carbachol silenced the raphe neurons almost completely. Rather than being a fundamental part of REM sleep or even an element in the mechanism that brings on an episode of REM sleep, the inactivity of the dorsal raphe neurons thus appears to be an ancillary phenomenon. It seems to be related to the central motor inhibition of REM sleep, which can be reversed by damage to the pons. Recently my student Peter Reiner has studied the other group of nerve cells that are turned off in REM sleep. These neurons lie near the site of our lesions and employ noradrenaline as a transmitter. Reiner's preliminary results suggest that the inactivity of these cells is also related to the motor inhibition of REM sleep.

By means of experimental techniques that enable us to separate the muscular paralysis associated with deep sleep from the activity that goes on in the brain we are learning much about what is fundamental to REM sleep and what is a side effect of REM-sleep episodes. The current hypotheses concerning the paralysis of REM sleep will undoubtedly require further work to be fully accepted, in particular the notion advanced above of separate pathways for the inhibition of muscle tone and of motor drive and the hypothesis that the second pathway is active in some way in wakefulness.

Whether or not these ideas prove to be sound, REM sleep without atonia



will continue to be a valuable experimental technique. In addition to making it possible to separate the essential from the ancillary in REM sleep, REM sleep without atonia could provide a means of studying complex behavior such as that of aggression, which is usually elicited in response to external stimuli but which could in REM sleep be generated in isolation from the animal's environment. Furthermore, because of the similarities between wakefulness and REM sleep it will undoubtedly continue to be interesting to compare the two states. In the movements of the waking animal we have a correlate of its inner states. Until now we have lacked such a correlate for REM sleep. In REM sleep without atonia we have acquired a correlate, which is in a sense a window on the sleeping brain.



SYSTEMS THAT INHIBIT MOVEMENT IN REM SLEEP may also operate in physical emergencies in wakefulness. The systems are shown in schematic form in the three panels of the illustration. Excitatory connections are shown in color and inhibitory connections in black. In normal REM sleep the pons strongly activates the inhibitory center in the medulla (*top*). The midline inhibitory zone in the pons inhibits the lateral locomotor strip. The result is complete paralysis. In REM sleep without paralysis the lesions break the connections from the pons to the locomotor strip and to the medullary center (*middle*). The cat can move, but the release of muscle tone is generally not complete. In wakefulness when an unexpected threatening stimulus is perceived, there is sometimes a transitory inhibition of movement (*bottom*). The temporary inhibition is probably the result of a reduction of motor drive, since there is little loss of muscle tone.

Early Farmers of the North European Plain

Excavations in Poland have revealed the remains of the people who farmed, herded and hunted in the area some 7,000 years ago. It was a time of transition, but in the end agriculture had come to stay

by Peter Bogucki and Ryszard Grygiel

One of the major events in human prehistory—the rise of agriculture and animal husbandry—took place at different times in different parts of the Old World. In southern Europe knowledge of this new way of life may have first diffused gradually from those parts of the Near East where the Neolithic Revolution had taken hold earlier. In Europe north of the Carpathians and the Alps, however, the Neolithic advance was a movement of people rather than ideas and one of the few such migrations clearly documented in the archaeological record.

Even so, after generations of field work not nearly enough is known about the people who left the valley of the middle Danube and carried the first cultivated plants, along with a number of domesticated animals, deep into the forested plain of northern Europe. The lack of data is due in part to the fact that the acid soils of the north quickly destroyed both human and animal remains and in part to the fact that archaeologists have preferred to work on larger and later Neolithic sites rather than on smaller and earlier ones. Nevertheless, in the 1930's early Neolithic sites on the Polish plain some 150 kilometers northwest of Warsaw had begun to yield valuable information until World War II halted the work. The same sites are continuing to yield information today. Before we describe them and their contents we should acquaint the reader with the physical nature of this broad transmontane belt and with the opportunities it offered to the first migrant farmers from the south.

Broadly considered, central Europe can be divided into three major ecological zones. The southernmost of them, the mountain areas of the Carpathians and the Alps, were generally not settled by farming communities until well into Neolithic times. Therefore that zone will not be discussed here. To the north beyond the mountains is a zone of rolling uplands extending from Slovakia and southern Poland through central Germany and into France. The region is largely covered with loess, the fertile but dry blanket of soil deposited by the winds that blew across Europe late in the Ice Age. The loess fills the river basins of the region and blankets the hills between them, reaching a depth of 100 meters in some places.

The third and northernmost zone consists of the flat lowlands of the north European plain, extending from the western U.S.S.R. across Poland, northern Germany and Denmark on into the Netherlands. The ice sheets of the last ice age covered this zone until some 15,000 years ago; their advance scraped out deep finger lakes and their retreat left a terrain of marshes, lake-filled hollows and a soil cover of sands, clays and gravels. Some of these new soils came to be very fertile; others remained marginal in fertility. All of them, however, supported a diverse and productive forest cover, and the lakes and streams were rich in aquatic life.

The farmers who entered the uplands from the south at first moved northwest along the major rivers of central Europe: the Danube, the Elbe, the Main and the Rhine. They settled in the richest upland ecological regions, the loess-filled river basins, where wild plants and animals were nonetheless relatively scarce. That left the newly established Neolithic communities largely dependent for their livelihood on their own imported animals (cattle, sheep, goats and pigs) and plants (barley and wheat). In time, however, some of the immigrants pushed farther north along the Vistula and Oder rivers and out onto the north European plain.

There they encountered terrain and soils considerably different from those of the uplands. The new ecosystem held more than 450 species of edible plants, fishes, turtles and mussels in the lakes and streams and red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), wild cattle, wild pigs and even wild horses in the forest. This zone of tremendous natural productivity had only one drawback. Its yield was not constant throughout the year and was much reduced in the winter months. How the early Neolithic settlers adjusted to this unevenly productive ecosystem is one major question for which the Polish sites are yielding an answer.

Of a total of more than 20 Neolithic sites now known in the vicinity of Brześć Kujawski, a town in the Włocławek district of Poland, two (designated Site No. 3 and Site No. 4) were discovered in 1933 when farmers digging gravel near the shoreline of a dry lake bed came on artifacts and human bones. Konrad Jażdżewski of the State Archaeological Museum in Warsaw learned of the discovery and spent seven seasons (1933–39) at the sites, excavating a total area of more than 10,000 square meters. His work exposed the outlines of many Neolithic houses, more than 50 human burials, some of them richly furnished with copper, bone and bead ornaments, and antler pickaxes. Both human and animal bones were excellently preserved in the nonacidic soil. The copper artifacts were some of the earliest to be found in that part of central Europe; the metal had probably come from mines south of the Carpathians. The war halted Jażdżewski's research, and the artifacts uncovered in the 1939 season were destroyed when the Warsaw rail depot was bombed, but his work had already established the importance of the sites to an understanding of early Neolithic life on the north European plain.

Both Jażdżewski's work, done in the years before carbon-14 dating was known, and later research in the area left a number of chronological and economic questions unanswered. As a result one of us (Grygiel), of the Łódź Museum of

OUTLINE OF LONGHOUSE WALLS, exposed by the removal of plow-zone topsoil, guided the excavators at one early Neolithic site at Brześć Kujawski as they sectioned the discolored clay soil where the timber uprights had stood. Similar subsoil stains indicated the presence of trash pits filled with discarded artifacts and animal bones. The irregular excavation in the foreground led to a bed of clay the Neolithic settlers used to "plaster" their house walls and make pottery. This is the later of the two main settlements, established by people of the Lengyel culture. Calibrated carbon-14 dates indicate they came here between 4500 and 4400 B.C. and abandoned the site in 3900 B.C.

THREE BURIALS were associated with the longhouse remains seen in the top photograph. Skeletons are those of a man, at the left, lying on his right side, and two women, lying on their left side. The orientation of the bodies at burial placed the heads in a south-southeasterly direction. More than 70 Lengyel burials have been excavated but no graves of the first inhabitants of sites No. 3 and No. 4, people of the Linear Pottery culture, have been found so far. Many Lengyel burials are accompanied by rich grave goods, including copper ornaments.

Archaeology and Ethnography, accompanied by the other (Bogucki), then a graduate student at Harvard University, reconnoitered sites No. 3 and No. 4 in the summer of 1976. We found that substantial areas of both sites were still unexcavated. Receiving financial grants for further work from the Włocławek district and the town of Brześć Kujawski, we have now finished seven seasons of excavation, mainly at Site No. 4.

As at most Neolithic sites in temperate Europe, the archaeological remains here lie at a relatively shallow depth below the surface. The topsoil is a rich medium-weight Kujavian "black earth" that rests on a clay and gravel substrate. The upper 30 centimeters are regularly disturbed by plowing each year, and so the method of excavation is to strip away the plow zone, thereby exposing any features dug into the sterile clay and gravel underneath. Unlike the layered accumulation of deposits typical of a deeply stratified site, what comes into view is a palimpsest of pits, wall trenches, postholes and graves, which frequently overlapped. Such overlaps make it possible to determine the relative ages of successive settlements. The absolute ages can be determined by carbon-14 measurements; a total of 15 have now been made by laboratories in Poland, the Netherlands and the U.S.

Two main periods of Neolithic occupation are evident at the sites. They are separated by an interval of two to three centuries when settlement was on a much reduced scale. The initial period began in about 5300 B.C. (the true calendar date according to calibrated carbon-14 chronology) and continued to about 4800 B.C. The occupants belonged to a group assigned by European archaeologists to the Linear Pottery culture because of the distinctive incised-line decorations of their wares.

DEER-ANTLER SPOON, unearthed from one of the many trash pits at Site No. 4, is the first such artifact to be found among central European Neolithic remains. The Lengyel people also made pickaxes out of the antlers of red deer they collected after the animals had shed them.

MUSSEL-SHELL BEADS, strung in strands, were among the grave goods found with this Lengyel woman's skeleton. One such burial included more than 5,000 one-centimeter shell beads.

The remains of the culture are found all across Europe from the Ukraine to France, both on the loess uplands and on parts of the north European plain. Many of the upland sites show traces of sturdy longhouses, sometimes only a single "farmstead" but more commonly several dwellings grouped into a village. No Linear Pottery longhouses have been found on the north European plain, however, and the two sites we excavated are no exception. Among the large but shallow trash pits that indicated the Linear Pottery presence were only a few scattered postholes, probably marking the location of lean-tos or similar temporary shelters. Two separate phases of Linear Pottery occupation could be perceived. The earlier was at the tip of a lakeshore peninsula at Site No. 3 and the later was about 200 meters away at Site No. 4. No Linear Pottery burials were found at either site.

Centuries later emigrants assigned by European archaeologists to the Lengyel culture arrived to occupy both sites. This early Neolithic manifestation, which is named after the original site in Hungary, represents an essential continuation of Linear Pottery traditions in east to central Europe but is characterized by new pottery forms and a decline in the amount of decoration on the pots. Many Lengyel sites also have longhouses that are trapezoidal in plan, being significantly narrower at one end than at the other. Longhouses of this distinctive form have been found both in the loess uplands and on the north European plain; they are the first permanent shelters to appear at the two Brześć Kujawski sites.

Four Lengyel phases can be identified at the sites. The earliest phase, evident mainly at Site No. 4, appeared between 4500 and 4400 B.C. and continued until about 4300 B.C. The three later phases, which extended from about 4300 until about 3900, are evident at both sites. In the 1930's Jażdżewski named these later local forms of the Lengyel culture the Brześć Kujawski Group, a name that is still in use.

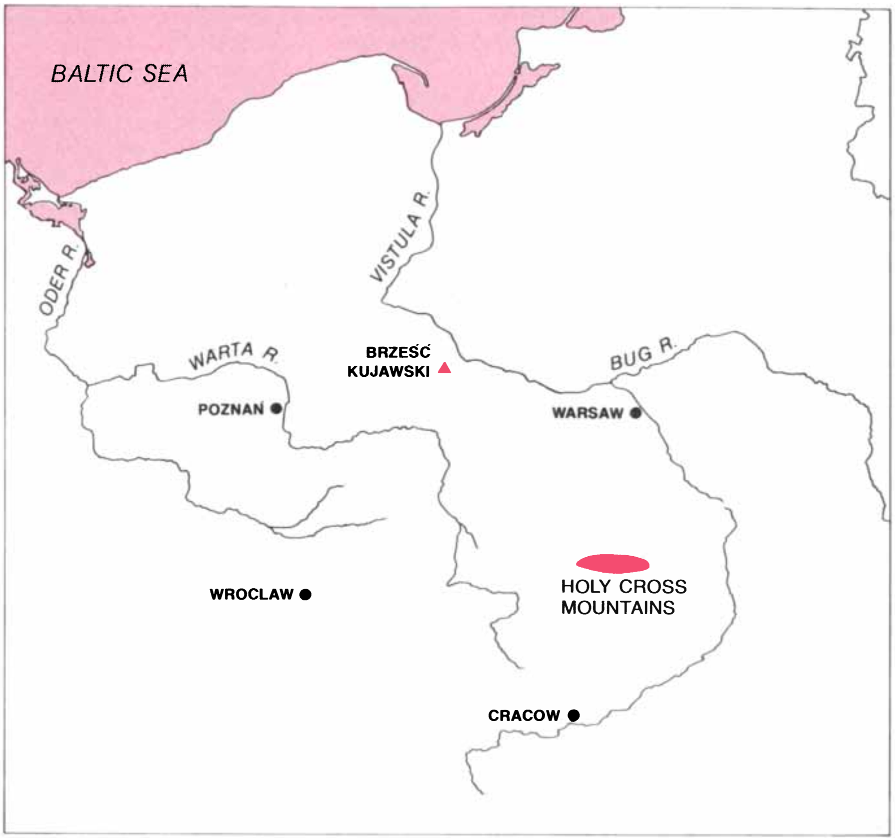
The most striking of all the features dug into the subsoil of the two sites are the traces of the Lengyel longhouses. They appear as long, dark discolorations in the clay and are the remains of trenches where the upright posts of the longhouse walls were set in place. Between the start of work in the 1930's and the present more than 50 such "house plans" have been uncovered, some of them more than 30 meters in length. All are oriented with their long axis running from northwest (the narrow end) to southeast (the wide end). On the average the Lengyel longhouses were some 20 meters long; the northwest end was about three meters wide and the southeast end about five meters.

The reason for the Lengyel people's building trapezoidal dwellings remains uncertain, but a probable explanation is that the narrow end faced into the prevailing winds. In 1976 a replica of a symmetrical Neolithic longhouse was erected in France as a television prop. The end facing the prevailing winds has already suffered severely from weathering but the other three walls are still sound. Several instances of Lengyel house rebuilding are evident at the two sites; the fact that the same locations were often reused suggests the occupation was continuous through all four phases, probably with as many as 10 houses occupied at any one time.

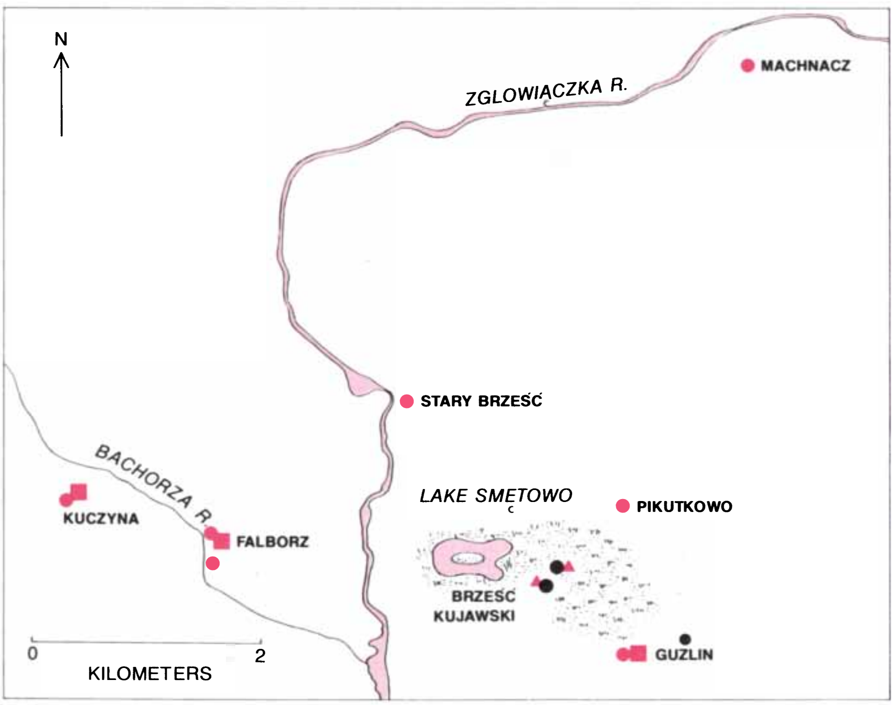
The commonest and most informative subsurface features at the two sites are pits of various sizes and shapes. As we have noted, the pits associated with the Linear Pottery people's trash are generally shallow and large. At the time of occupation they may well have been nothing more than natural depressions in the ground, such as one left when a falling tree is uprooted. They contain dense accumulations of debris, consisting of animal bones and discarded stone, flint and pottery artifacts. In contrast, many of the Lengyel pits are deep, irregular holes dug in parts of the site where clay lies close to the surface. This suggests that the Lengyel people originally dug them to get clay for "plastering" their timber dwellings and for making pottery.

Other Lengyel pits are generally deep and circular and are found near the shore of the former lake. Silty deposits at the bottom of the pits suggest that at times they held standing water, probably for the maintenance of live mussels, turtles and fish. In any event most of the clay-yielding pits and the maintenance pits eventually served for trash disposal. Unlike the Linear Pottery pits, the Lengyel pits show considerable variation in their trash content. Animal bones and discarded artifacts are more numerous in the pits nearest the longhouses.

In some areas where the subsoil was disturbed in Neolithic times the discoloration is less than it is in the pits filled with trash; apparently the area that had been dug up had been refilled without any admixture of organic debris. On investigation such areas usually proved to be Lengyel graves. More than 70 Lengyel burials have now been excavated, the majority of them by Jażdżewski in the 1930's. The bodies were usually found in this kind of individual grave pit. Generally the skeletons are in a contracted position with the knees drawn up; males were interred lying on their



BRZEŚĆ KUJAWSKI SITES, indicated by a colored triangle, are 150 kilometers west-northwest of Warsaw and about 200 kilometers northwest of the Holy Cross Mountains, an area where chocolate flint, named for its distinctive color, is found. This unique flint was prized by the Linear Pottery people because of its excellent flaking properties. Their successors at Brześć Kujawski were largely content with glacially deposited local flint of poorer quality.



SUCCESSIVE OCCUPATIONS at Brześć Kujawski stood on a peninsula jutting out into what was then a much larger Lake Smętowo. The Linear Pottery occupations at sites No. 3 and No. 4 were near the tip of the peninsula and near its base (*black dots*); the Lengyel occupations, at the same two sites, appear as colored triangles. Other Linear Pottery sites (*colored dots*) have been found to the south, west and north. Three nonresidential Lengyel sites (*colored squares*) have also been found. One of them, judging by the many axes found there, was a lumber camp.

right side and females lying on their left. With both sexes the horizontal orientation is the same: the skull rests near the south to the southeast edge of the pit. The females were often adorned with beads and other ornaments of shell, bone and copper; antler pickaxes are frequently found with the males. So far no Linear Pottery burials are known.

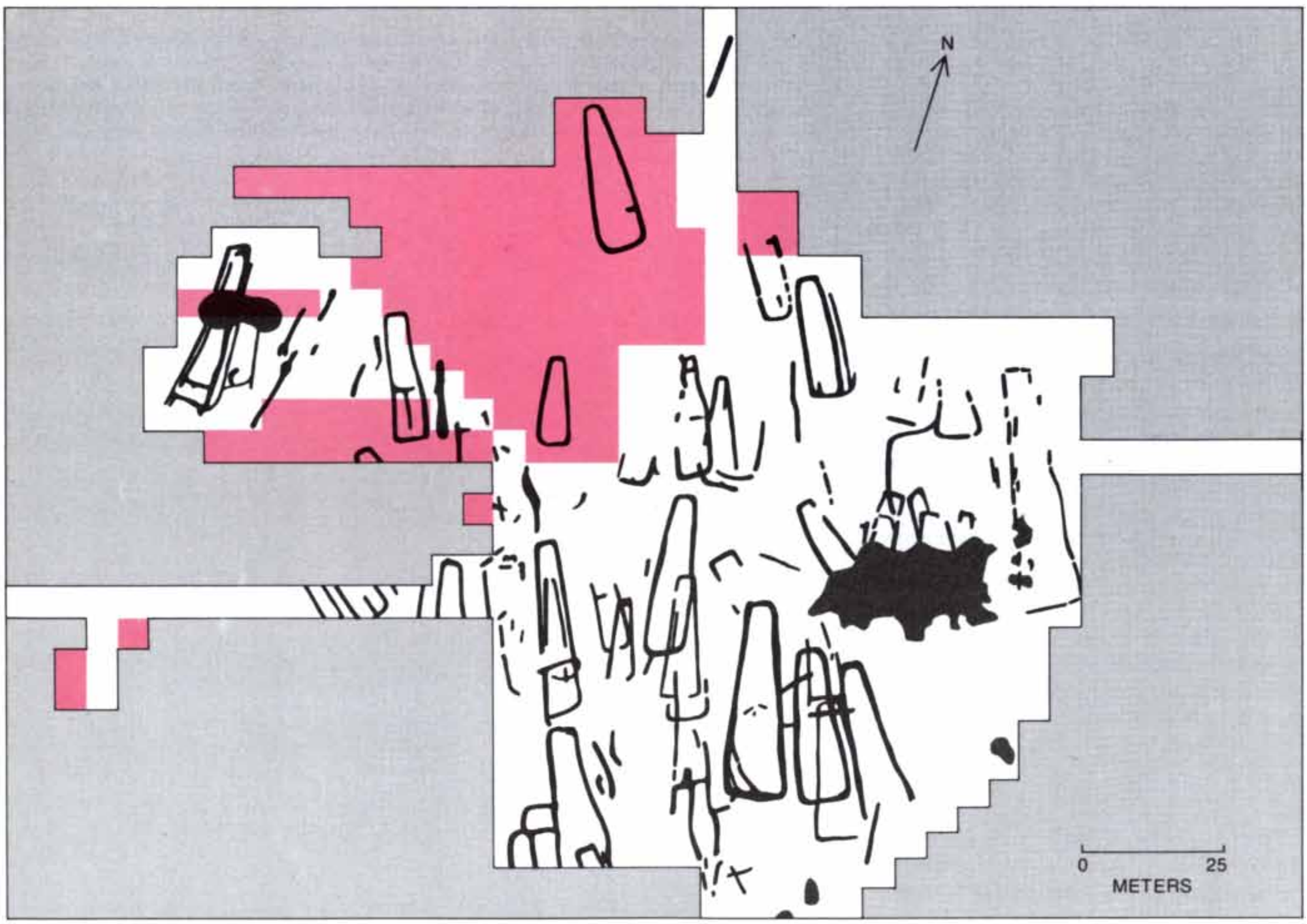
The assemblages of artifacts left by both the Lengyel and the Linear Pottery peoples include characteristically Neolithic axes and adzes made out of "pecked" and polished stone. Indeed, an ax of this type (known as the shoe-last celt) is characteristic of the Linear Pottery culture. At the same time these early farmers still made various small tools by flaking flint. The flintwork of the two cultures, however, was quite different. The Linear Pottery people favored a Polish specialty: chocolate flint, so called because of its brown color and found only in the Holy Cross Mountains some 200 kilometers to the southeast of the sites. Polish chocolate flint was valued by the makers of stone tools even in Paleolithic times. It has excellent flaking qualities, and the Linear Pottery people struck long, thin blades from chocolate-flint cores.

The Lengyel people showed no such strong preference. They made most of their flake tools from the locally plentiful flint "erratics" that the retreating ice sheet had left liberally scattered across the north European plain. Flint erratics, many of which have been subjected to pressure and other stresses in the course of being transported by the ice sheet, tend to have relatively poor flaking qualities, and the Lengyel flint flakes show it. Only when the Lengyel people needed a sharp precision tool did they take the trouble to procure and flake chocolate flint.

The Lengyel people went a step beyond the Linear Pottery people in making pecked and polished stone axes by drilling them so that a handle could be inserted; the stone plugs that were removed in the procedure are found in the trash pits. The Lengyel people also collected the shed antlers of red deer both to make pickaxes and to use as the raw material for other artifacts. Most notable among them is a handsome antler spoon, unique in central Europe, that we uncovered only last year.

Animal bones and teeth were also fashioned into beads and pendants. The ribs of large mammals (probably cattle) were bent into armlets, apparently while still "green," and then decoratively incised. Huge quantities of beads were made out of the pearly shell of the freshwater mussel (*Unio*). Strands wrapped around the waist of a female in one burial consisted of more than 5,000 *Unio* beads, each one about a centimeter in diameter. The graves of Lengyel females also contain copper beads and amulets; strips and bars of copper discarded in the trash pits indicate that these ornaments were made on the spot out of imported copper "blanks."

The pottery after which the Linear Pottery culture is named differs from the Lengyel pottery not only in decoration but also in form. At Brześć Kujawski the commonest shape is the same as the one most frequently encountered at other Linear Pottery sites in central Europe: a deep bowl forming three-quarters of a sphere. Another Linear Pottery shape that is common at the two Polish sites and elsewhere on the north European plain but relatively rare in the uplands is a sieve. We found no



MORE THAN 50 TRACES OF LONGHOUSE WALLS, many of them superposed, were excavated by Konrad Jażdżewski of the State Archaeological Museum in Warsaw at Site No. 4 between 1933 and 1939. The house outlines visible in the white area of this plan show the work of the 1930's. The dark gray areas indicate modern disturbance before Jażdżewski began work. Colored areas were excavated by the authors between 1976 and 1982. To avoid clutter burials and trash pits are omitted from the plan (*see illustration on opposite page*).

intact sieves, but the number of sieve fragments suggests that these artifacts served to separate curds from whey in making cheese. Robert Cowie of the University of Sheffield is currently conducting studies of Linear Pottery potsherds to see if their mineral composition offers clues to the sources of the clay used in their manufacture.

The Lengyel wares have shapes ranging from shallow bowls to deep vessels rather like Greek amphoras and includes such specialized items as pottery spoons. Their mineral content has not been subjected to detailed analysis, but the large amount of mica in the wares suggests they were made out of materials available locally. The same is suggested by the fact that clay was removed from pits at the site.

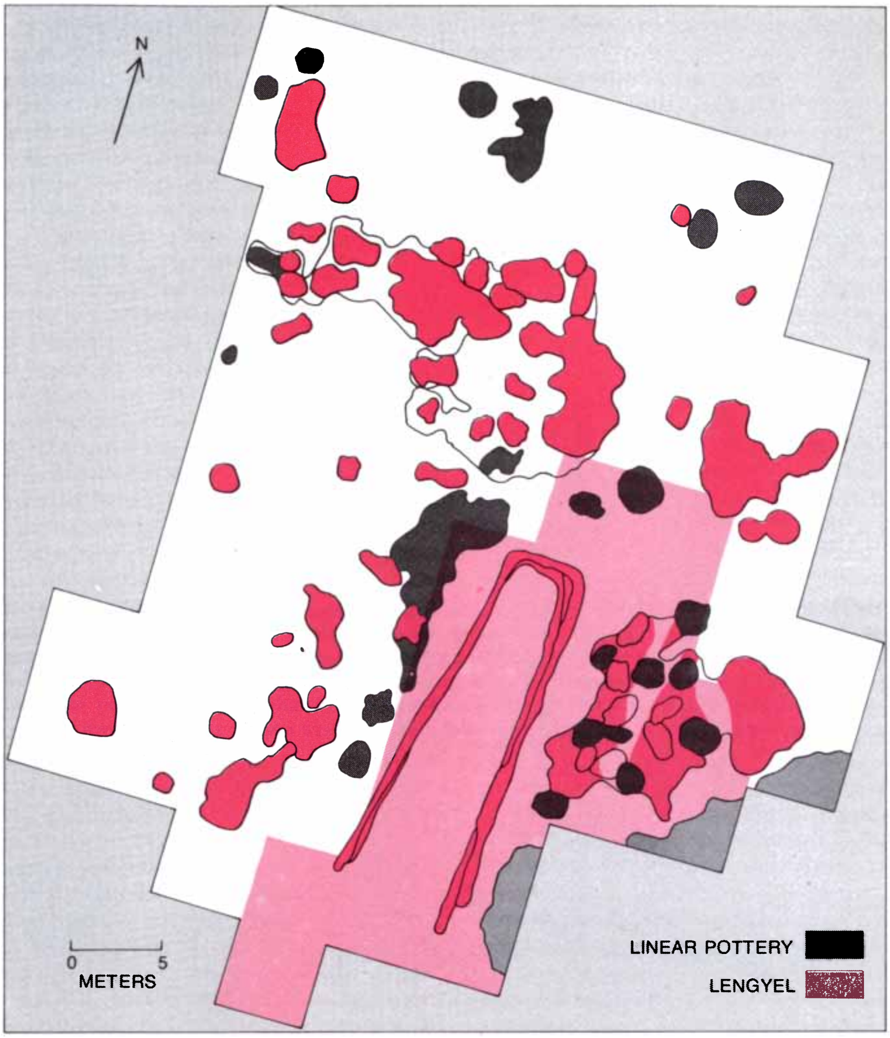
In our work since 1976 we have recovered a great many animal bones in a state of good to excellent preservation. For example, we have separated out more than 6,000 fishbones, mostly those of perch and bream. Here again there is a contrast between the animal remains characteristic of the earlier occupants of the site and those characteristic of the later occupants. As is the case with other sites of the culture on the north European plain, the primary domestic animals of the Linear Pottery culture were cattle. Since most of the cattle bones are at the small end of the overall size range, the herds appear to have consisted mainly of cows. Some sheep and goats were also tended by the Linear Pottery people, but whereas these farmers kept many pigs in their upland settlements, we have unearthed traces of only three pigs over the entire 500 years of the Linear Pottery occupation of Brześć Kujawski. Also notable for their small numbers are the remains of birds, fish, turtles and mussels and the bones of game animals: red deer, roe deer and wild pigs. Indeed, of the large number of fishbones found at the two sites, fewer than 100 come from the Linear Pottery occupation.

The animal bones associated with the Lengyel occupation tally differently. Although overall the remains of Lengyel cattle are somewhat more numerous than those of Linear Pottery cattle, these animals do not figure as prominently in the livestock totals. Sheep and goats are almost as numerous, and the pigs equal the cattle in numbers. Simultaneously there came an increase in the harvest of wild game: red deer and roe deer, waterfowl, turtles, fish and mussels. Not all the hunting was for meat. For example, beavers were evidently killed for their pelts, and so too, perhaps, were otters, weasels and hares.

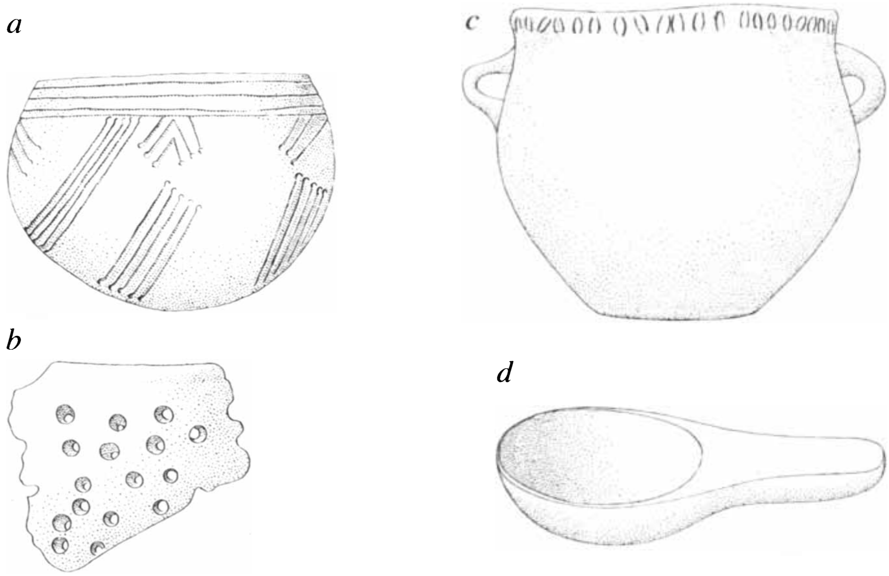
What about the tending of crops at these early Neolithic settlements? We have paid particular attention to the recovery of plant materials by means of wet-sieving and flotation techniques. We have succeeded in finding only three kernels of carbonized grain in the Linear Pottery trash pits, which suggests that these people engaged in very little grain cultivation. The Lengyel trash pits, in contrast, have yielded copious amounts of carbonized grain. One pit we excavated last year yielded more than 300 specimens. Preliminary analysis by Caroline Quillian Stubbs, who was then working at the University of Tübingen, indicates that most of the Lengyel grain is the early form of wheat known to botanists as emmer (*Triticum dicoccum*).

It is not clear whether the Lengyel people cleared forest to enlarge their farming activities. Such a conclusion would call for the analysis of pollen grains at various levels in the soil. No such pollen profiles are yet available from our two sites, although work to this end should soon be finished. One possible indication of forest-clearing activities, however, is the presence of the bones of eagles and marsh hawks among the animal remains. Both birds require open areas as part of their habitat.

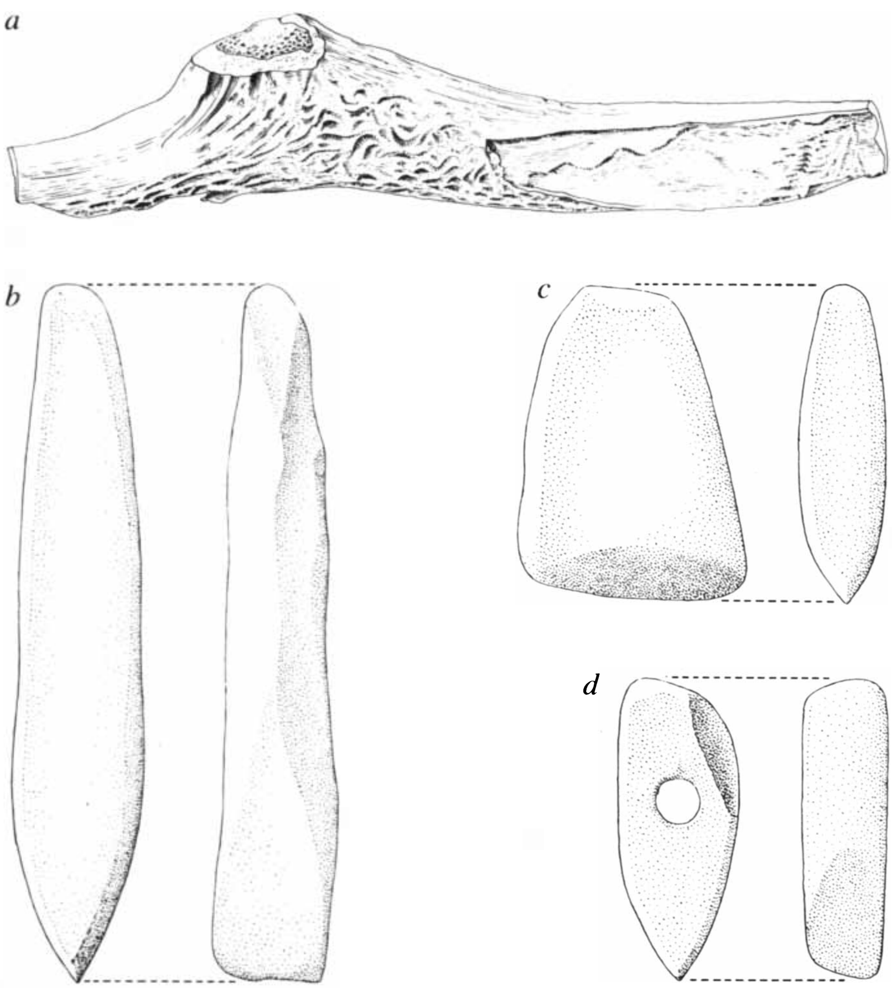
To determine what resources the early Neolithic populations exploited at various times of the year, we depended on a number of different clues, including the teeth of pigs and deer, and the degree to which the teeth still in the jawbones of the animals had erupted. We also noted other indications of seasonality such as the presence among the animal bones of the remains of certain waterfowl species (indicative of winter and spring hunting), of pond-turtle remains (the animals are inactive from late fall to early spring) and of shed deer antlers (they are cast annually in February and March). We could find no indication of winter activities in the Linear Pottery trash. At the same time the presence of turtle



WORK AT SITE NO. 3, nearer the lakeshore, done by Jażdżewski lies within the colored area of this plan. It included the exposure of two longhouse outlines almost perfectly superposed and a number of similarly overlapping trash pits that contained artifacts and animal remains of both the Linear Pottery and the Lengyel occupants of the site. The trapezoidal shape of the Lengyel longhouses is evident; the narrower end may have faced into the prevailing wind. Although the longhouses of the Linear Pottery people are found elsewhere in Europe, none are known on the north European plain. Postholes associated with Linear Pottery artifacts here suggest that these earliest Neolithic migrants constructed shelters of only a temporary kind.



DEEP POTTERY BOWL shaped like a three-quarter sphere (*a*) displays the linear style of decoration that gives the Linear Pottery culture its name. Below it (*b*) is part of a sieve, a common Linear Pottery artifact at the Polish sites, although intact sieves have not been found. The sieves probably served to separate curds from whey in making cheese. The simply decorated pot with ears (*c*) is a typical Lengyel form. Below it (*d*) is another common clay object of Lengyel manufacture: a pottery spoon. Neither the sieve sherd nor the spoon is shown to scale.



AXES were made out of pecked and polished stone or red-deer antlers (*a*). The stone ax at the left (*b*), known as a shoe-last celt, is a characteristic tool of the Linear Pottery people. The shorter, broader ax (*c*) is a characteristic Lengyel form. The Lengyel people also drilled holes for handles in some of their axes (*d*). No intact drilled axes have been found at sites No. 3 and No. 4, but the stone plugs removed in making the holes have been found in trash pits.

bones in it indicates that the Linear Pottery group was active at the sites in the summer months. In contrast the contents of Lengyel trash deposits demonstrate that these later occupants of the sites lived there the year round.

The examination of "growth rings" in the teeth of mammals proved to be particularly informative. These layers of cementum are deposited as opaque bands in summer and as translucent bands in winter. By sectioning a tooth and determining which kind of layer was last deposited one can broadly determine when in the year the animal died. In this way we found that the Lengyel people usually slaughtered their pigs in midwinter and hunted red deer and roe deer in late winter and spring.

In addition to our work at sites No. 3 and No. 4 we reconnoitered the nearby countryside in search of other early Neolithic sites. We found evidence of six other Linear Pottery occupations, all showing the same characteristics as the main sites near Brześć Kujawski: they lacked evidence of permanent dwellings and their rubbish pits were shallow. Three additional Lengyel sites have also been discovered, all within an hour's walk of the two main sites. They also lack evidence of any permanent dwellings and the few pits they contain generally yielded arrays of artifacts more limited in variety than those at the main sites. They give the impression of being areas where some specialized activity took place, perhaps crop tending or herding. One of them, at Kuczyna, appears to have been a lumbering camp: its trash pits yielded an inordinate number of stone and antler axheads.

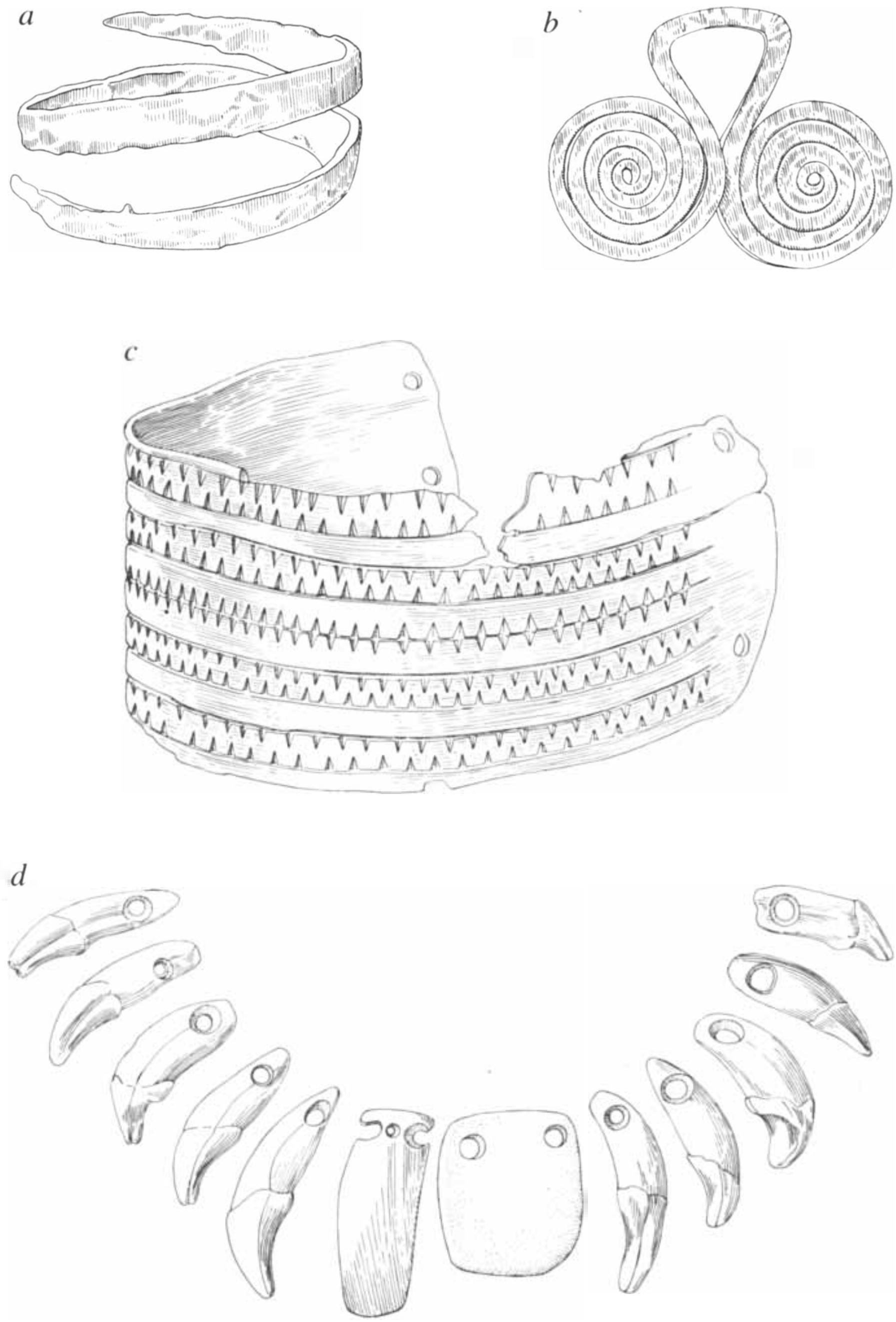
In all, both the nearness of the Lengyel camps to the main settlement area and their lack of permanent structures suggest that the Lengyel farmers occupied a well-defined "territory," whereas the more extended scatter of Linear Pottery camps gives the opposite impression. This, together with the functional differentiation of the Lengyel satellite camps, indicates that the Lengyel farmers exploited their surroundings in a more organized fashion than their Linear Pottery predecessors.

What are the general conclusions to be drawn from the work at Brześć Kujawski? In the past both the subsistence economy and the material culture of the first Neolithic communities in central Europe have often been implicitly characterized as "package deals." For example, wherever Linear Pottery is found it is presumed that an entire "package" of longhouses, pecked and polished "shoe last" axheads, grain, cattle, sheep, goats and pigs will also be present. At the two Polish sites, however, such is not the case, and neither is it at other Linear Pottery sites on the north European plain. The upland longhouses are absent, the evidence of grain cultivation is sparse and pig bones are a rarity. Moreover, the Linear Pottery occupations of sites No. 3 and No. 4 show no indications of winter residence. The opposite is true of the Lengyel occupations of the same sites and also of Linear Pottery sites in the upland zone to the south.

We interpret this information as calling for a reconsideration of the way of life in the earliest food-producing communities of the north European plain. In view of the omissions from the "package" and the high proportion of cattle bones among the Linear Pottery animal remains we suggest that this first Neolithic excursion north of the uplands consisted not of permanent farming settlements but of seasonal cattle herding (and cheese making) in the lowland forests. The ecologically diverse vegetation of the lowlands would have provided rich summer forage for cattle in spite of the timber cover, probably far richer than the grazing available in the narrow valley bottoms of the uplands.

The conditions favorable for seasonal cattle herding may have been further enhanced in this forest area by an activity of the indigenous Mesolithic (pre-Neolithic) inhabitants of the area: the deliberate clearing of standing timber by fire. Evidence that such was a Mesolithic practice is provided by the abnormally high proportion of hazel pollen in the pollen profiles representing the period dated at about 5800 B.C., some 500 years before the first Linear Pottery people entered the lowlands. Hazel is a common second growth on cleared land.

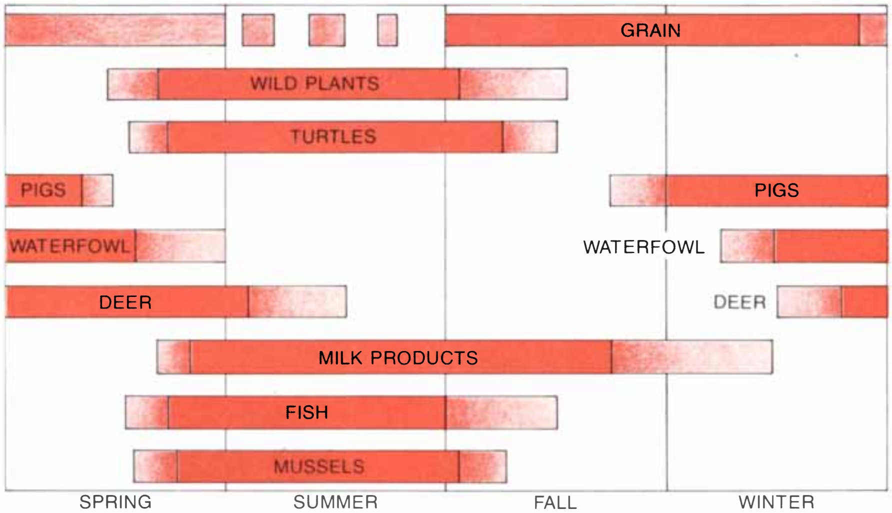
Added evidence of the importance of dairy products to these earliest Neolithic immigrants is the presence of clay sieves both at Brześć Kujawski and at other Linear Pottery sites on the north European plain; these artifacts are comparatively rare at upland Linear Pottery sites. One may presume that one reason for it is that the upland settlers had not only grain but also numerous pigs, sheep and goats, whereas the lowland diet was limited to wild plant foods and dairy products. Another reason may be that the Linear Pottery people of Neolithic central Europe had not yet developed the tolerance to lactose (milk sugar) characteristic of modern European peoples; therefore they needed to process cows' milk into cheese (and perhaps yogurt) in order to consume dairy products without indigestion. Given the entire array of evidence, it seems probable that Linear Pottery herders came north with their cattle in the warm season but returned to permanent settlements in the uplands for the winter.



TYPICAL ORNAMENTS from Lengyel burials are a copper bracelet (*a*) and a copper "spectacle" pendant (*b*). Incised bone armlets (*c*) were made out of animal (probably cattle) ribs. The necklace (*d*) combines polished fossil-coral pendants with dogs' or wolves' teeth drilled for stringing. The copper artifacts, surprising in an early Neolithic context, were probably made on the spot out of copper carried here from mining areas south of the Carpathians.

The Lengyel farmers who succeeded the Linear Pottery people on the lowlands left behind evidence of the kind of diversification of resources that might be expected of a newly established year-round agricultural economy. Although they grew grain, it is doubtful that the emmer harvest was large enough to feed the community until the next harvest and at the same time provide seed for the spring planting. They must certainly have collected such wild produce as hazelnuts and goosefoot (*Chenopodium*). Goosefoot is a source of both greens and seeds that was widely exploited elsewhere in Neolithic Europe. It grows to a height of 1.5 meters in this part of Poland. The Lengyel people may also have supplemented their grain crop with such garden produce as peas and lentils; both are found at other Neolithic sites in central Europe, although no evidence of either has yet turned up at Brześć Kujawski.

Of even greater importance in the Lengyel diet were the domestic animals the settlers herded and the wild game they hunted. As our tooth studies affirm, the Lengyel people slaughtered their pigs in winter after a summer of fattening, and they hunted red deer and roe deer for meat in late winter and spring. Turtles, fish and mussels would also have contributed to their diet in spring and summer, supplementing the wild

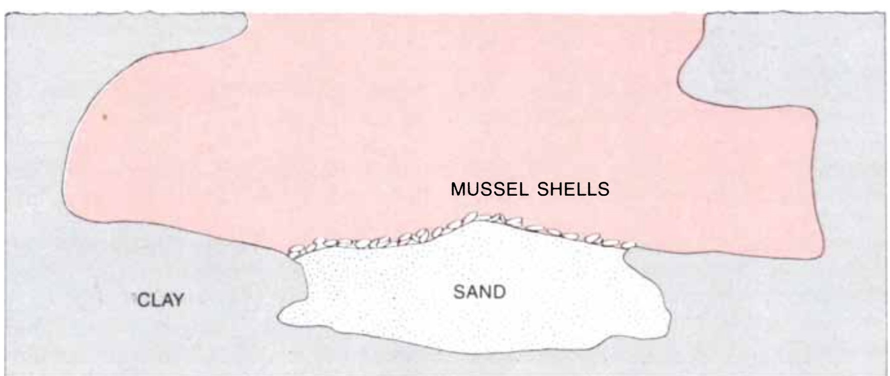


YEARLY FOOD CYCLE for the Lengyel people, beginning at the end of winter (*left*), saw a dependence on one domestic animal (the pig) and three wild animals (migratory waterfowl, red deer and roe deer) to bridge the gap until wild plant foods, freshwater mussels, turtles and fishes, together with milk from cattle, sheep and goats, became available in late spring and on into fall. With the wheat harvest in the fall the Lengyel people acquired a major resource that saw them well into the winter, when the summer-fattened pigs again became an important source of food, supplemented by waterfowl and deer and by cheese, a storable dairy product.

plant foods from the surrounding forest and the milk provided by cattle, sheep and goats. Such a combination of wild and domestic foodstuffs would have been particularly welcome in late summer, when the reaping and storing of the emmer crop would have called for a great deal of work in a short time.

The organizing and scheduling of this kind of subsistence system was evidently what made necessary a change in Lengyel times from the dispersed settlement pattern of the Linear Pottery people. We have mentioned the Lengyel satellite camps that contributed to the maintenance of the permanent residential area; there were probably also a number of other locales—hunting stations and herding outposts—that have left no trace in the archaeological record. The surrounding woods, fields and marshes must have been known intimately by the Lengyel farmers, every square meter contributing something to the well-being of the central settlement.

The long continuity of the Lengyel occupation testifies to the success with which these Neolithic farmers exploited the forest ecosystem around them. Any ecosystem, however, is resilient only up to a point. It seems likely that gradual environmental deterioration after hundreds of years of human occupation was what forced the abandonment of the Brześć Kujawski sites by about 3900 B.C. By then, however, food production as a way of life had become permanently established on the north European plain. The Neolithic Revolution was spreading along the Baltic coast and into neighboring northern Germany and Denmark in the hands of still later Neolithic peoples such as the people of the Funnel Beaker culture. The Linear Pottery and Lengyel pioneers were no more, but agriculture, with its innumerable consequences, had come to stay in temperate Europe.



PIT AT SITE NO. 3, close to the former lakeshore, is shown in simplified cross section. The presence of intrusive sand and a thin layer of mussel shells suggests that the pit was a "storage well," with standing lake water filling its bottom, used to keep mussels, turtles and fish alive.

TAKE A LOAD OFF YOUR MIND WITH A FREE FINANCIAL PLANNING GUIDE FROM IDS.

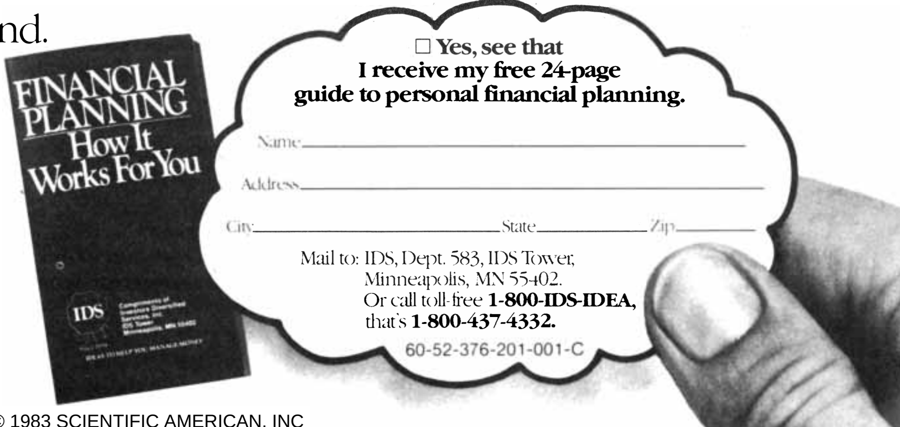
Financial worries can spoil your appetite in a lot of ways. Interest rates fluctuate, tax laws change and inflation hasn't disappeared. Banks, brokers and insurance companies all seem to give conflicting advice. It's confusing. What you need is a carefully thought-out, completely objective plan that's flexible enough to work in the years ahead. And that's where IDS comes in. We're offering you a free guide to financial planning that explains what a financial plan is and how it can diminish your money worries.

An IDS representative can help you develop a personal financial plan to achieve your financial goals. You'll be talking to a career professional who understands the importance of financial planning. Someone who will regularly review your progress.

And because IDS offers over 50 different financial products and services, you'll get timely ideas to help you reach your goals.

Begin taking a load off your mind. Send for your free 24-page guide to personal financial planning. Or call toll-free 1-800-437-4332.

IDS IDEAS TO HELP YOU MANAGE MONEY



Chemical Signals of Social Amoebae

Two cellular slime molds can coexist in the same soil and yet maintain their identity. They do so by emitting and responding to different chemicals, both of which have now been identified

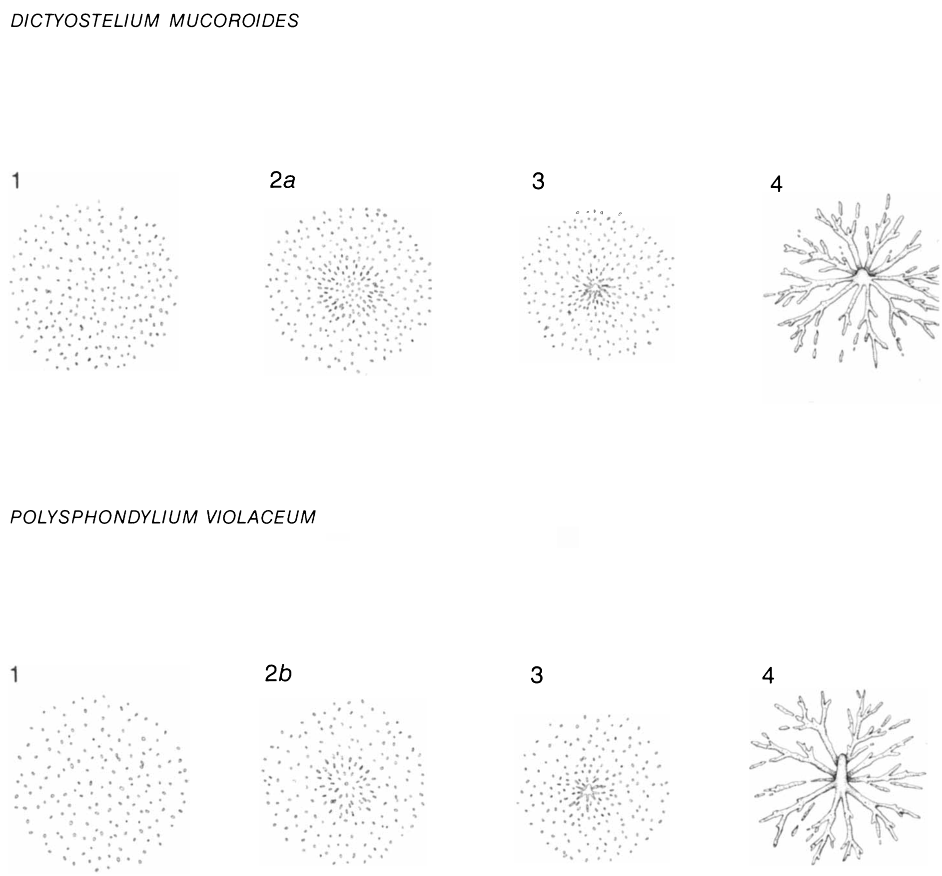
by John Tyler Bonner

In any bit of reasonably fertile soil two very small unicellular organisms are abundant: bacteria and amoebae. There may be millions of bacteria in a cubic centimeter of soil, and thousands of amoebae. The bacteria tend to be clumped near dead roots or dead insects, which supply them with nutriment, and the amoebae eat the bacteria. Each amoeba is about the size of a white blood cell, some 10 times longer than a bacterial cell. It crawls through the soil, engulfing bacteria it encounters, and every few hours it divides to form two daughter cells.

Prominent among the commonest kinds of soil amoebae are the social amoebae, otherwise known as the cellular slime molds. They feed as separate cells, like other amoebae, but once they have finished off all the bacteria in an area a trigger is set off within each cell that initiates their social phase. The cells stop their solitary wandering and begin to stream toward a central collecting point. Anywhere from a few hundred to 100,000 or so unicellular organisms come together to create a single, sluglike mass that has become a multicellular organism.

Then the cells begin to differentiate. A nipplelike tip appears at the front end of the mass, and inside the tip the beginning of a stalk takes shape. The stalk is a cellulose cylinder composed of swollen, rigid cells that ultimately die. The stalk builds upward as cells stream up the outside of it and pile onto the tip. When all movement has ceased, the mass of cells has become a fruiting body: a simple or branched stalk made up of dead cells encased in a cellulose sheath and capped by a little ball of spores. The spores are individually encapsulated amoebae, which become dispersed by water or by contact with a passing worm or insect. If the spore lands in a suitably warm and moist environment, it splits open, releasing a single amoeba that resumes the foraging phase of the slime-mold life cycle.

What I want to discuss in this article is a particular aspect of the social amoebae's behavior. Often there are two or more species of social amoebae in the same bit of soil. When they aggregate, they aggregate separately: the cells of different species become oriented toward separate collection centers. The streams of cells heading for one center pass through, and ignore, streams headed for another center, so that each species comes together at its own collection center and differentiates to form its characteristic fruiting body. How does it happen? Pieces of the answer have been emerging for some 40 years, and now it is finally possible to describe the segregation of the two species at the molecular level.



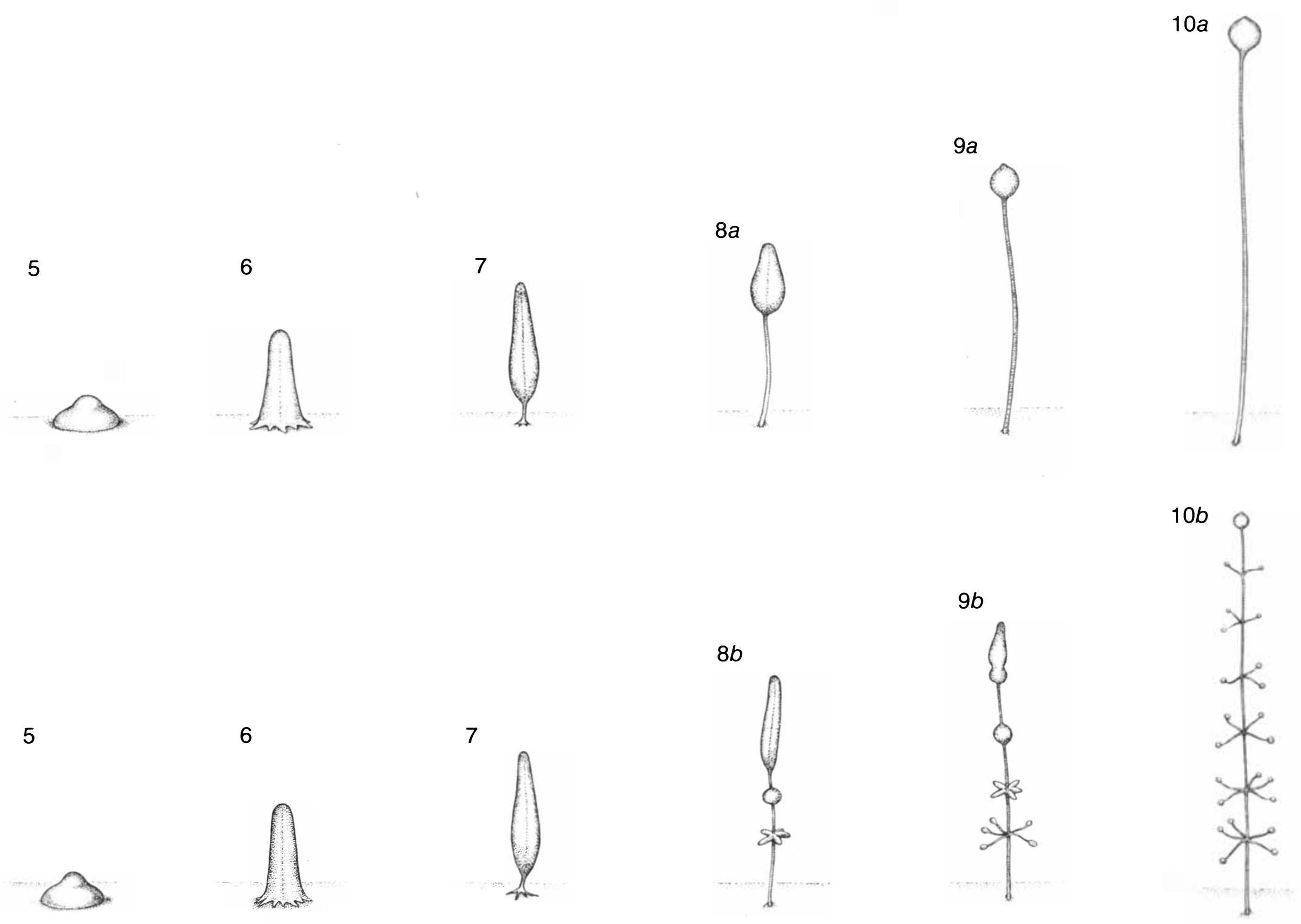
LIFE CYCLE AND MORPHOLOGY of two cellular slime molds, *Dictyostelium mucoroides* and *Polysphondylium violaceum,* are similar. They feed as individual amoebae (*1*) until lack of food (bacteria) triggers either a group of cells (*2a*) or a single cell (*2b*) to secrete a chemical substance, called an acrasin, that attracts surrounding amoebae (*3*). The cells stream toward a

The segregation phenomenon was noted in 1940 by Kenneth B. Raper and Charles Thom, who were then working in the U.S. Department of Agriculture. They grew cultures of the bacterium *Escherichia coli* and inoculated a culture dish with two species of social amoebae, *Dictyostelium discoideum* and *Polysphondylium violaceum*. The amoebae multiplied and grazed their way outward from the point where each species had been inoculated; eventually the expanding populations were thoroughly mixed. Having consumed the available bacteria, the amoebae in the region of mixing began to aggregate. As Raper and Thom wrote, "they regularly aggregated to different centers with the converging streams of the two forms commonly overlapping.... It was obvious that the stimuli in different species differed qualitatively." Raper and Thom recognized why that had to be. If the two species are together in the same soil, they must have a way to avoid coming together in the same aggregate in order to remain distinct species. Raper and Thom could not know how the two stimuli differed because the nature of the stimuli was not then known.

At the turn of the century scientists studying cellular slime molds had assumed that the amoebae must be brought together by some form of chemotaxis, but in the 1940's accounting for a developmental process by chemical attraction was viewed with suspicion; some form of "contact guidance" was a much more popular explanation. In 1942 Ernest H. Runyon of Agnes Scott College made it clear, however, that contact among cells could not bring about aggregation. In a series of experiments over the next few years I was able to show that the central mass had to be emitting some diffusible attracting agent. It could have been heat, but a number of indications made it clear that the agent must be a chemical substance. I called the substance an acrasin. (Acrasia is the name of a witch in Edmund Spenser's *Faerie Queene* who attracted men and transformed them into beasts, and *Dictyostelium* is a member of the Order Acrasiales.)

Over the years studies in numerous laboratories, including our own at Princeton University, have revealed many details of the acrasin system. Starvation is the stimulus for the switch from the feeding state to the social one. It sets in motion a series of chemical events within the individual cells, which soon begin to secrete an acrasin. At the same time the amoebae synthesize and display on their surface a large number of specific protein molecules that serve as receptors for the acrasin. The cells also secrete an acrasinase (an enzyme that inactivates the acrasin) and an inhibitor of the acrasinase. The acrasin, the acrasinase and the inhibitor combine to form a complex control system that modulates aggregation as it proceeds. To take just one example, the acrasinase keeps the external acrasin level generated by an individual amoeba low enough so that the cell can sense acrasin emitted by other cells.

Aggregation begins differently in different species, as Brian M. Shaffer, who was then working at the University of Cambridge, found some years ago. In



central collection point and pile up to form a multicellular organism (*4*). A tip forms on the cell mass (*5*) and the cells begin to differentiate. Some of them form a rigid internal stalk (*6, 7*). The remainder are borne aloft on the stalk to become spores. In *Dictyostelium* there is a single spherical mass of spores at the tip of a mature fruiting body (*8a, 9a, 10a*). In *Polysphondylium* there are whorls of small fruiting bodies in addition to an apical spore mass on the central stalk (*8b, 9b, 10b*). The spores become dispersed and split open, freeing amoebae.

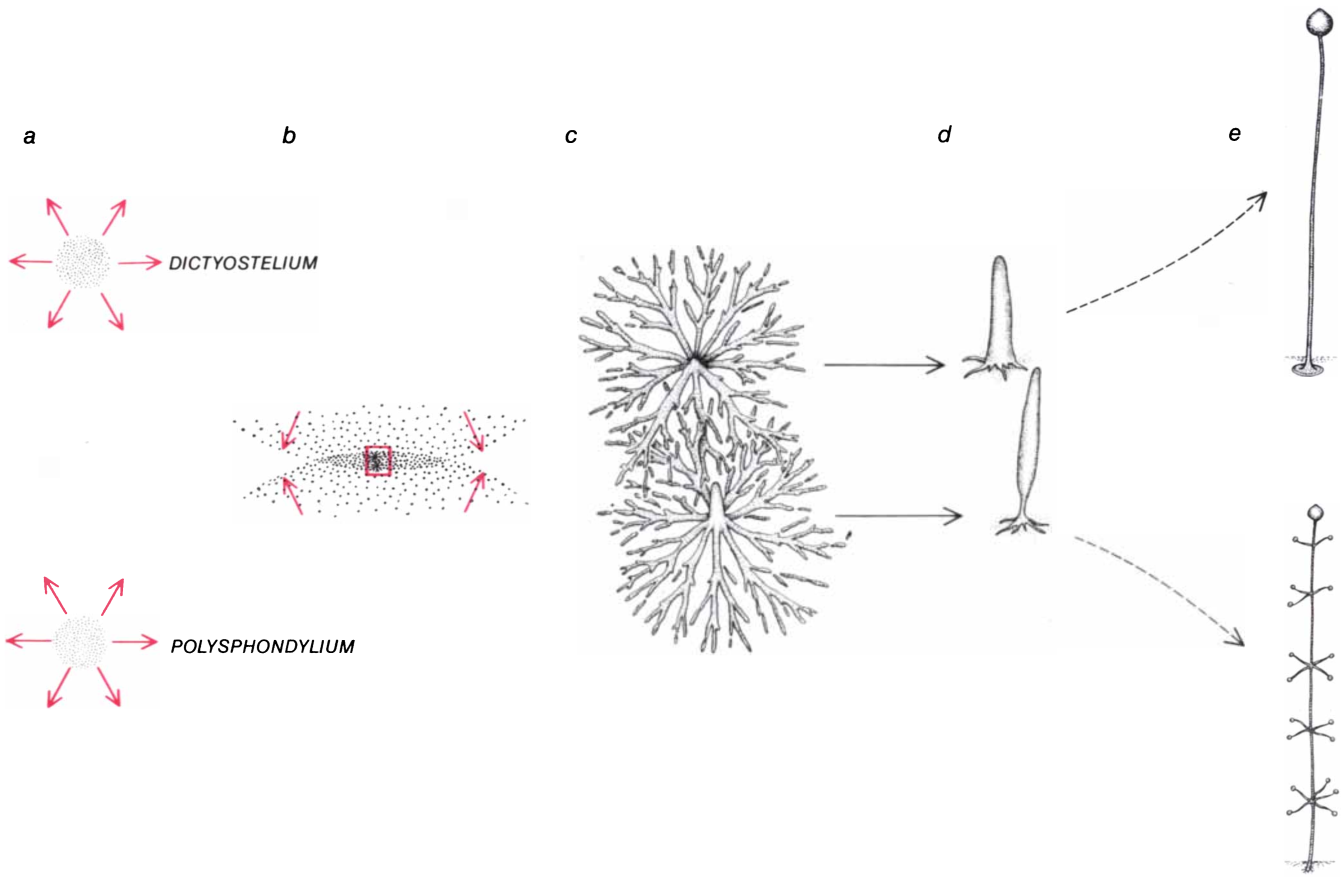
Polysphondylium a single cell in the population suddenly rounds up and begins to secrete acrasin. Neighboring cells streak in to join it, oriented by the concentration gradient of the secreted acrasin. In *Dictyostelium,* on the other hand, a small group of cells (Shaffer called it a cloud) seems to initiate aggregation. These cells first become relatively immobile and then clump together, and the clump attracts the surrounding cells. In both species the aggregating amoebae often come inward in pulses. There seems to be a good reason: to establish an effective overall gradient of acrasin that will orient distant cells, a good-sized central mass is required. The pulses can orient cells before such a gradient is established. If one amoeba gives off a puff of acrasin and the puff diffuses out to neighboring cells, it hits each cell on one side first, filling more of the cell's receptors on that side than on other sides, and thereby orients the cell. The responding amoeba in turn emits a puff that attracts cells beyond it.

When chemotaxis was firmly established as the mechanism of aggregation, I argued at first that it was not important to know the chemical nature of the acrasins. I say this with some embarrassment because the fact is that the moment the chemical structure of a key substance in physiology or development is discovered, one goes from the dark ages to modern times. Certainly that was true in the case of the *Dictyostelium* acrasin. In any event we soon joined other laboratories in an effort to characterize its attractant. It took from 1947 to 1967. I have told the story in detail before [see "Hormones in Social Amoebae and Mammals," by John Tyler Bonner; SCIENTIFIC AMERICAN, June, 1969], and I shall only summarize it here.

The major difficulty was developing a convenient bioassay: a way to test the biological effect of successively purified substances. Such assays were finally developed by Theo M. Konijn (who is now at the University of Leiden) and by us. We both did preliminary work toward identifying a factor found in bacteria that seemed to attract amoebae. Konijn came to Princeton, and soon after he arrived David S. Barkley, who was then a graduate student, had an inspiration: Why not try cyclic adenosine monophosphate (cyclic AMP), a nucleotide that Earl W. Sutherland, Jr., of the Vanderbilt University School of Medicine had discovered and had shown to act in mammalian cells as a "second messenger," mediating between hormones and cell functions.

It worked, with remarkable potency, in our chemotaxis assays. We went on to show that *Dictyostelium* cells not only respond to a minute amount of cyclic AMP but also produce it. (Now, to be sure, it is well known that almost all cells make cyclic AMP; not producing it is the rarity.) More to the point, we found that during aggregation the amoebae secrete the nucleotide in particularly large quantities (a hundredfold increase) and also are about 100 times more sensitive to the cyclic AMP than they are at other times.

The discovery that cyclic AMP is the natural acrasin of *Dictyostelium* opened up new lines of investigation. Workers in laboratories all over the world have studied just how the nucleotide orients cells and have analyzed the activity of adenyl cyclase, the enzyme that catalyzes the synthesis of cyclic AMP, and of phosphodiesterase, the acrasinase for this acrasin; they have examined the



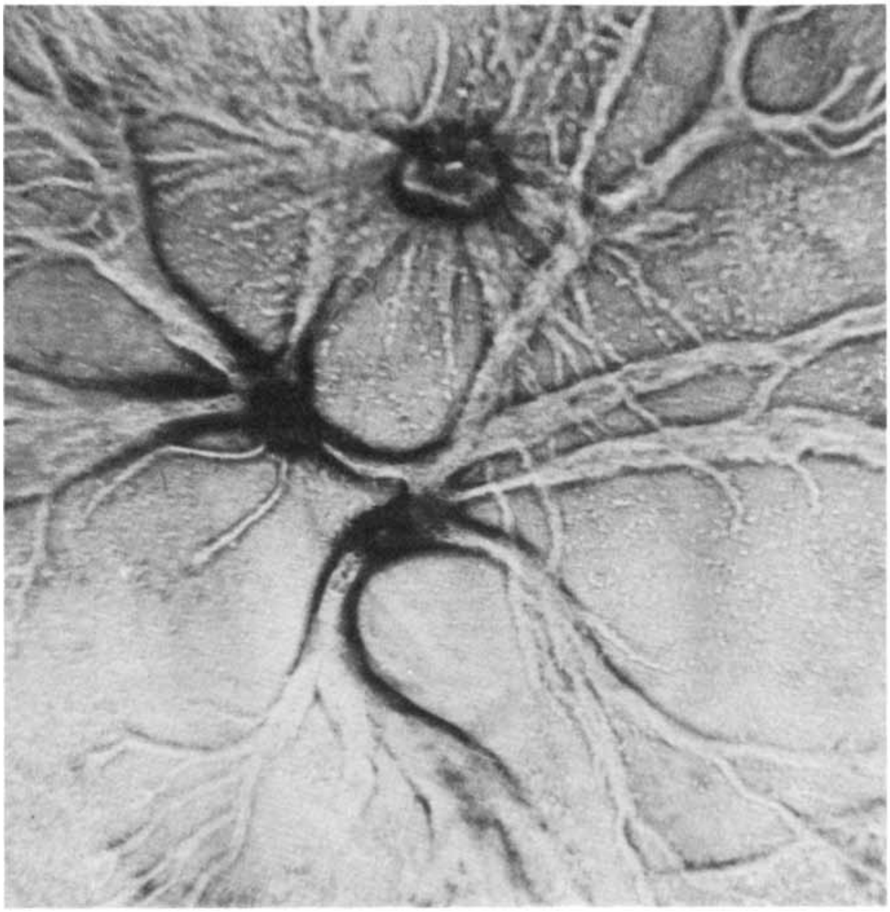
TWO SLIME-MOLD SPECIES aggregate separately, as Kenneth B. Raper and Charles Thom demonstrated in 1940 with this experiment. A culture of the bacterium *Escherichia coli* is inoculated with individual cells of *Dictyostelium discoideum* (*top*) and *P. violaceum* (*bottom*). The amoebae graze outward from the point of inoculation (*a*), become mixed and begin to aggregate (*b*). When the aggregations are examined under the microscope (*c*), they are seen to be separate. Streams of amoebae of the two species head for two different collection centers and form separate organisms (*d*), which develop to form distinctive fruiting bodies that are characteristic of the two species (*e*).

cell-surface receptors for cyclic AMP and the interactions of all the elements of the system. In 15 years there has been great progress toward understanding how some aspects of the development of *Dictyostelium* are controlled, and the new knowledge has implications not only for the development of the cellular slime molds but also for development in general.

As soon as we knew that cyclic AMP attracts *Dictyostelium* cells we tried it on *Polysphondylium*. It had no effect, just as Raper and Thom might have predicted (and as Shaffer later had predicted). The two species have different acrasins. The ensuing effort to identify the *Polysphondylium* acrasin was very different from our *Dictyostelium* experience. There was no clever guess by a bright graduate student. We could not even find a good source of unpurified attractant; bacterial products and even human urine had been good sources of cyclic AMP. We tried a variety of things that are generally rich in hormones and other active substances—milk, peptones (mixtures of peptides), yeast extracts and many others—with no effect. We had no recourse but to set about extracting the material from aggregating amoebae. The moral is that one cannot be blessed with fantastic luck every time.

The first step was to find out how best to extract the acrasin from an aggregation of *Polysphondylium*. As we had learned working on *Dictyostelium*, the acrasinase secreted along with acrasin tends to break down the attractant as soon as it is secreted. We finally settled on a method devised by David W. Francis, who was then a postdoctoral fellow at Princeton and is now at the University of Delaware. It consists of dumping 40 percent ethyl alcohol directly on petri dishes containing cells in midaggregation, washing the dishes with more alcohol and then collecting the wash in a beaker. We boil off the alcohol and centrifuge off the precipitate that forms. A crude extract of acrasin is left in solution; it remains intact because the alcohol has denatured the acrasinase.

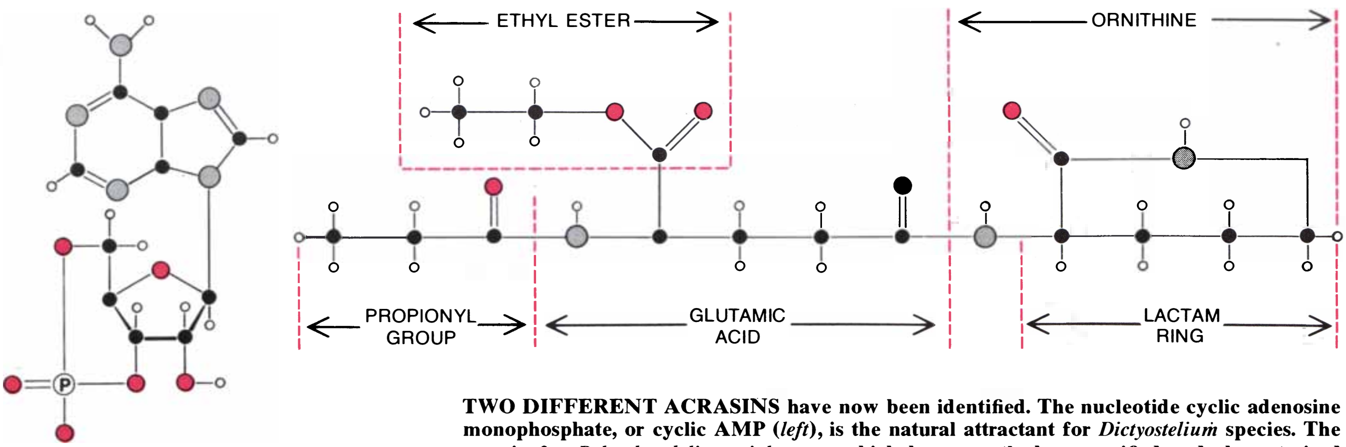
The remarkable thing is that the amoebae do not seem to be bothered by this 80-proof dunking; they go right on with their development after the alcohol has been removed. In order to collect as much crude extract of *Polysphondylium* acrasin as possible, over the years we have developed a routine that begins with 150 large petri dishes of aggregating amoebae, which cover every available surface in the laboratory. It has the disadvantage of stopping all other experiments for lack of space, but the laboratory does have a rather pleasant aroma. At first we used to think we could smell the acrasin, but it turned out to be the alcohol.



SEPARATE AGGREGATION of two species was recorded in this photomicrograph, made by Raper and Thom, in which the streams of aggregating amoebae are enlarged some 45 diameters. There is a small aggregation of *D. discoideum* (*top*) and a larger aggregation, with two dark centers, of *P. violaceum*. The inflowing streams of amoebae overlap, with the cells of each species apparently ignoring those of the other species as they head for their own center.

In the early days of this project, almost a decade ago, we sought the major characteristics of the *Polysphondylium* acrasin. This was done by Bernd Wurster (who is now at the University of Konstanz in Germany) and Pauline Pan (now with the Pennwalt Corporation). After more than two years of work we could say that the acrasin was a small molecule, with a molecular weight of less than 1,500 (estimated from its movement on a gel filtration column); that it was stable when heated; that it did not appear to be ionic in nature (that is, that the molecule was uncharged), and that it did not have any free amino or carboxyl groups, which are usually found respectively at the beginning and the end of a protein chain. Yet two enzymes—one protease and one peptidase—of a number we tried completely inactivated the acrasin, suggesting it was some kind of peptide (a short piece of a protein chain). Unfortunately this was not a conclusive finding because the enzyme preparations were impure; something quite secondary could have been playing a role. These were, however, neutral enzymes, which attack uncharged polypeptides, and this was consistent with the observation that the acrasin molecule seemed not to be charged. We concluded that the *Polysphondylium* acrasin might be a peptide in which there was no free amino group, with some evidence that it incorporated what is called an ester bond.

Further analysis was difficult because in spite of our best efforts our purest extracts were very impure. It was at this point that our colleague Osamu Shimomura, who is now at the Marine Biological Laboratory in Woods Hole, Mass., became interested and took charge of all the chemistry, including the purifications. To give him enough material to work with we had to increase the number of harvests. Under the direction of Hannah B. Suthers we started an "acrasin factory": two days of the week were devoted entirely to harvesting the 150 large petri dishes of acrasin. This went on over three years, for a total of 96 harvests. In the first two years we were still experimenting, perfecting the method of purification. In the third year we made what turned out to be the final push to accumulate crude acrasin, and

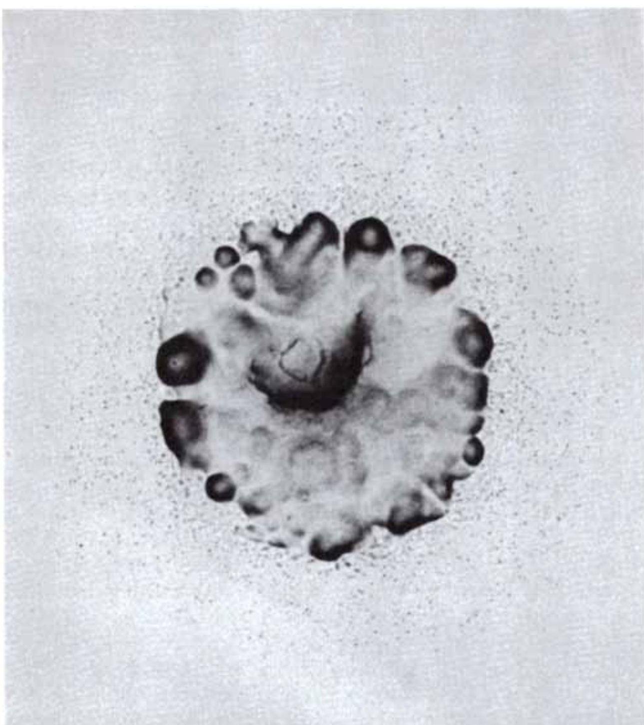


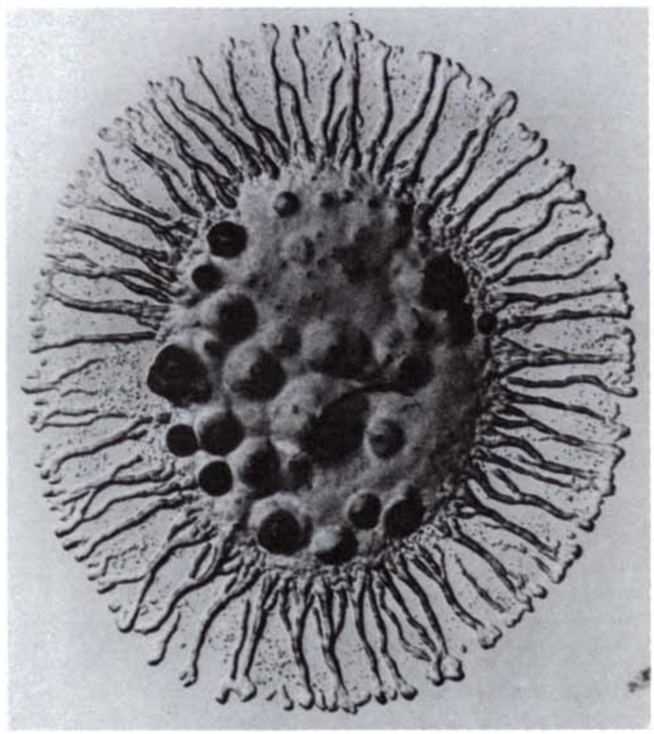
TWO DIFFERENT ACRASINS have now been identified. The nucleotide cyclic adenosine monophosphate, or cyclic AMP (*left*), is the natural attractant for *Dictyostelium* species. The acrasin for *Polysphondylium violaceum,* which has recently been purified and characterized by the author and his colleagues, is a dipeptide that has been named glorin (*right*). Its major components are the amino acids glutamic acid and ornithine. An amino group (NH_3) and a carboxyl group (COOH) of the glutamic acid are blocked respectively by a propionyl group and an ethyl ester. An amino group on the ornithine molecule is blocked by a lactam ring. Both cyclic AMP and glorin are small molecules, with molecular weights of 324 and 327 respectively.

40 collections gave us an initial four grams of very impure material. Shimomura put this through a series of nine purification steps involving three different kinds of gel filtration columns and numerous solvent systems. At the end we had 92 micrograms (millionths of a gram) of acrasin, about 98 percent pure.

The 92 micrograms of acrasin was divided into lots of about 10 micrograms that were sent away for different kinds of analysis. From the infrared spectrum we learned that there was indeed the strong possibility of an ester group. Amino acid analysis confirmed an earlier finding that the molecule contained two amino acids, glutamic acid and ornithine (which is not one of the usual components of proteins), in roughly equal amounts. The excitement came from the results obtained by Catherine E. Costello of the mass-spectrometry facility at the Massachusetts Institute of Technology. Using the latest magic (ionization techniques of field desorption and fast-atom bombardment, at both low and high resolution), it was possible in one final push to know the complete structural formula and the molecular weight. This is quite remarkable, considering that it comes from looking at a sample of about 10 micrograms of material.

The *Polysphondylium* acrasin turned out to be a dipeptide. There is a glutamic acid molecule with two side groups: a propionyl group, which blocks the amino end of the amino acid, and an eth-



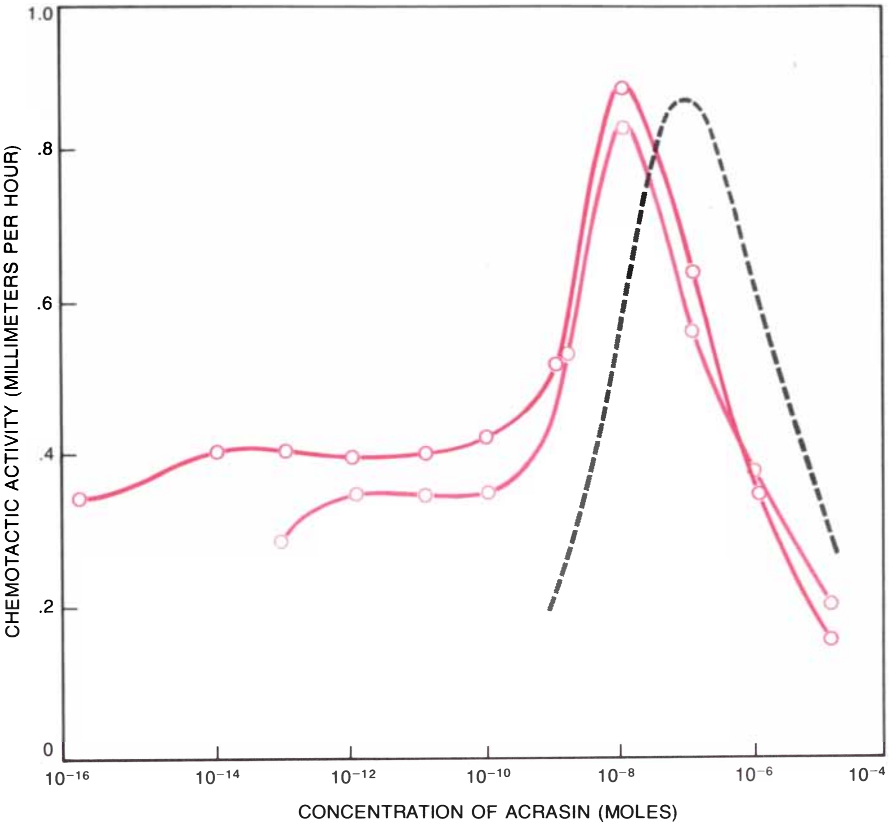


BIOASSAY demonstrates that glorin is indeed the natural attractant for *P. violaceum.* When a large clump of amoebae is placed on a block of agar (*left*), all but a few of the cells stay in the clump. When a clump of amoebae is placed on a block of agar that has been impregnated with synthetic glorin (*right*), the amoebae stream outward vigorously, attracted by the glorin. (The glorin directly under the clump is rapidly destroyed by an acrasinase, an inactivating enzyme secreted by the cells. *P. violaceum*'s acrasinase has not yet been identified.)

yl ester, which blocks the free carboxyl group. The glutamic acid is attached by a peptide bond to an ornithine, but the ornithine is closed into a lactam ring, which again blocks a free amino group. The molecular weight of the acrasin is 327. Because it consists of glutamic acid and ornithine, whose abbreviations are *Glu* and *Orn,* we decided to call the acrasin glorin.

To confirm that glorin is the naturally occurring acrasin of *Polysphondylium* we needed to have the dipeptide synthesized and then test the manmade version in our chemotaxis assay with *Polysphondylium* amoebae. We sent our glorin formula to Peninsular Laboratories, Inc., and after a month or so they delivered to us 50 milligrams of synthetic glorin. (We calculated that it would have taken more than 500 years to accumulate that much material by extraction and purification in our laboratory!) At M.I.T., Costello immediately showed that its mass-spectral properties were identical with those of the natural, purified glorin. At Princeton, with considerable anxiety, we undertook the bioassay.

The result was unequivocal. The synthetic glorin had enormous biological activity: when it was placed in agar surrounding a mass of test amoebae, they streamed in a most dramatic way [*see bottom illustration on opposite page*]. In a quantitative test we compared the synthetic glorin's ability to attract amoebae with that of the natural, purified glorin (the remains of our precious hoard) and found the two were essentially identical in their activity. When we compared the glorin curves with data showing the effect of cyclic AMP on aggregating *Dictyostelium* amoebae, it was clear that glorin works effectively at even more dilute concentrations than cyclic AMP does.



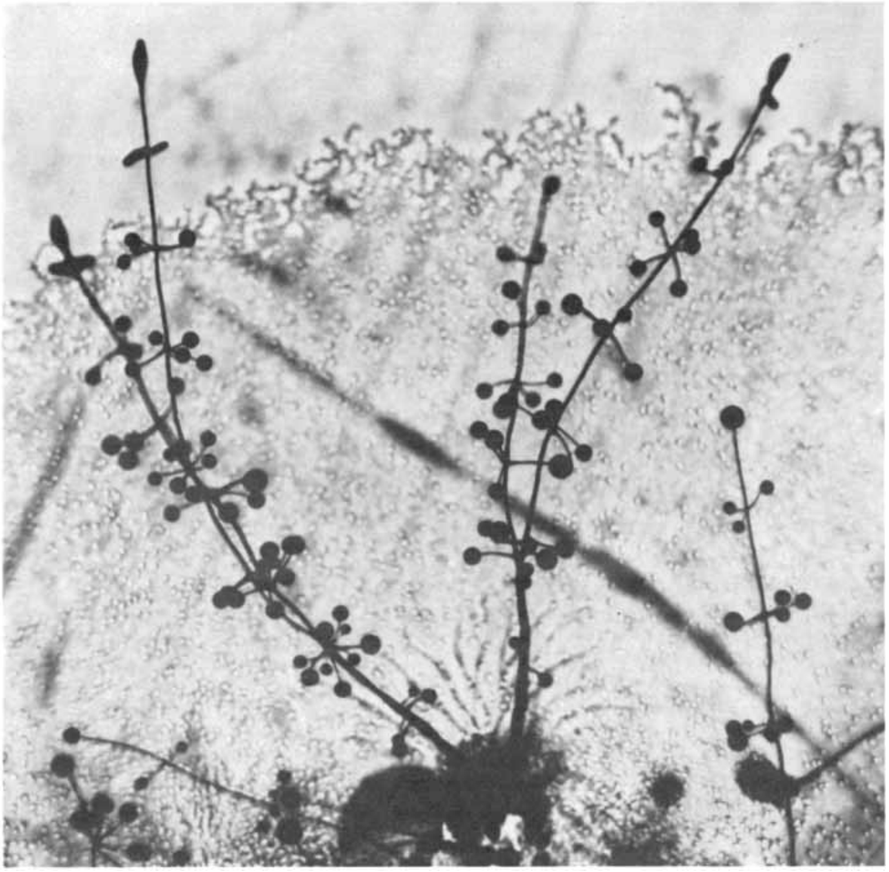
CHEMOTACTIC ACTIVITY of glorin synthesized in the laboratory (*dark color*) is compared with the activity of the natural attractant (*light color*) purified from aggregating cells. The two curves are essentially identical, confirming the chemical analysis of the natural glorin. A similar curve (*gray*) shows the effect of cyclic AMP on aggregating cells of *D. discoideum*.

So much for the excitement of the chase. What can we now do with glorin? One thing is to work out all the details of the glorin aggregation system, including the identification of its receptors, the acrasinase and the inhibitor of the acrasinase. We also want to ask this question: What are the respective contributions of glorin and of cyclic AMP to differentiation? The reason for this interest is that is has been known since our early experiments with Konijn that cyclic AMP is present in *Polysphondylium.* Later, in our laboratory and in others, mounting evidence was found that cyclic AMP is involved in the differentiation of both stalk cells and spores in *Dictyostelium,* and there is similar evidence in *Polysphondylium.* There is even interesting work in *Dictyostelium* showing that cyclic AMP is directly involved in controlling the expression of some developmental genes. Recent work in laboratories in the Netherlands and the U.S. has provided convincing evidence that chemotaxis continues to operate in the cell mass during later stages of development: a cyclic AMP gradient within the multicellular organism orients the cells, thereby guiding the direction of the advancing tip. During these later stages of *Dictyostelium* development, then, cyclic AMP is involved in two things, chemotaxis and differentiation, at the same time and in the same place.

We want to find out if in *Polysphondylium* there is perhaps a separation of functions, with chemotaxis controlled by glorin and differentiation controlled by cyclic AMP. The relation of the two signaling systems may not be simple and separate; it is quite possible that they are mutually dependent on one another or that glorin is more directly involved in cell differentiation than we currently assume. Here lies a unique opportunity to compare two parallel developmental systems that have different chemical signaling systems. Among other things, it should provide a way to separate, isolate and identify the controls of the different developmental events.

Perhaps the main point is that *Dictyostelium* and *Polysphondylium* live side by side in the soil, are similar in appearance and go through remarkably similar stages of development. Yet they have totally different signaling systems, one system based on a nucleotide and the other on a peptide. The only comparable situation is found in the body of animals, including human beings, where cyclic AMP acts in many signaling reactions, both hormonal and neural, and a number of peptides function in the nervous system as neurotransmitters. The presence of rival signal systems based on small nucleotides and on peptides is a very general biological phenomenon, not just a slime-mold curiosity.

Some broad evolutionary questions arise from the fact that these two signaling systems are so different. I pointed out above that Raper and Thom's experiments seemed to imply that by responding to two different stimuli two species living in the same place can keep from intermingling and thereby losing their identity. The problem extends beyond two species. James C. Cavender of Ohio University and Raper, who have spent much time collecting slime molds all over the world, have shown that in some temperate and tropical environments as many as from five to eight species may coexist. It is not necessary for all of them to have different acrasins to keep separate. As Raper and Thom showed in their original paper, there can also be "cell-surface incompatibility": some



STALKS AND SPORE MASSES of *Polysphondylium pallidum* (which, like *P. violaceum*, secretes and responds to the acrasin glorin) are enlarged 16 diameters in a photomicrograph made by Raper. They have developed at the center of a bacterial colony, whose edge is seen at the top of the picture. The fruiting bodies are in different stages of development: the larger ones are still building and the one at the right, with a spherical apical spore mass, is mature.

species aggregate together, in response to the same acrasin, and then the cells of the two species separate from each other at the center of the aggregate to form two separate multicellular individuals. Nevertheless, there is already evidence that there are at least eight different acrasins among the approximately 50 known species of cellular slime molds; when all these species have been carefully investigated, additional acrasins will surely be revealed.

This means that the extent to which cellular slime molds have been able to evolve from one chemotaxis system to another is quite astounding. It raises the fascinating question of how a signaling system can shift from one such chemical system, with its acrasin, its receptors for a specific acrasin and no doubt other associated substances, to another system. Clearly the selection pressure on slime molds to diversify and occupy new niches must be intense; even given that assumption, how are the evolutionary steps of change from a cyclic AMP chemotaxis system to a glorin chemotaxis system (or vice versa) accomplished?

I find it useful to compare slime molds in the soil to signal systems in man. Human beings have a vast array of hormones and neurotransmitters, and one assumes they arose during man's evolution from simpler animals that had far fewer signal systems. This leads to the same question as before: How did new hormones and neurotransmitters arise during the early evolution of animals? Again one could imagine that for increased control it is important to have a series of quite separate hormones and neurotransmitters, and that through natural selection some of those signaling systems gave rise to new ones, just as slime molds evolved into new species.

I am drawing a grand analogy between a mammalian body and a plot of soil and postulating for both a trend toward the increased diversification of chemical signals. It is therefore conceivable that the origins of acrasins may help us to understand the origins of animal neural and hormonal signaling systems. There is, however, a difference of staggering proportion between slime molds and mammals: the signals in animals, including human beings, produce an individual organism with extraordinarily unified accomplishments, including noble thoughts, something one can hardly expect of a shovelful of soil.

Where can you get a front row seat to hear the National Symphony every week?

The Mutual Broadcasting System now brings the National Symphony as close as your radio. Under the musical direction of Mstislav Rostropovich, you'll hear a brilliant concert season, arranged just for you.

Once a week, on your Mutual radio station, you'll hear the classics you love—Mozart, Brahms, Beethoven and more. You'll hear performances from the Kennedy Center for the Performing Arts in Washington, D.C., and concerts from the Symphony's recent European Tour. Some of these concerts will even be coming to you live, via satellite, in stereo.

Check your local radio listings for the next performance. Then take the best seat in your house. Because for the best in classical music, the answer is Mutual, your full-service network.

The answer is Mutual.

MUTUAL BROADCASTING SYSTEM

Intuitive Physics

Although Newton's laws are well known, tests show many people believe moving objects behave otherwise. The subjects of the tests tend to follow a theory held in the three centuries before Newton

by Michael McCloskey

Imagine the following situation. A person is holding a stone at shoulder height while walking forward at a brisk pace. What will happen if the person drops the stone? What kind of path will the stone follow as it falls?

Many people to whom this problem is presented answer that the stone will fall straight down, striking the ground directly under the point where it was dropped. A few are even convinced that the falling stone will travel backward and land behind the point of its release. In reality the stone will move forward as it falls, landing a few feet ahead of the release point. Newtonian mechanics explains that when the stone is dropped, it continues to move forward at the same speed as the walking person, because (ignoring air resistance) no force is acting to change its horizontal velocity. As the stone travels forward it also moves down at a steadily increasing speed. The forward and downward motions combine in a path that closely approximates a parabola.

The discrepancy between intuition and physical reality in this situation is somewhat surprising. One might expect that as a result of everyday experience people would have reasonably accurate ideas about the motion of objects in familiar situations. One might even suppose people would have an informal grasp of the general principles that govern objects in motion. It seems that such is not the case. Recent studies on the nature, development and application of knowledge about motion indicate that many people have striking misconceptions about the motion of objects in apparently simple circumstances. The misconceptions appear to be grounded in a systematic, intuitive theory of motion that is inconsistent with fundamental principles of Newtonian mechanics. Curiously, the intuitive theory resembles a theory of mechanics that was widely held by philosophers in the three centuries before Newton.

In a series of studies done by my colleagues and me at Johns Hopkins University, high school and college students were asked to predict how a moving object would behave in diverse situations. One experiment concerned the motion of an object that is initially confined to a circular path and then is suddenly released. Such motion takes place, for example, when an athlete throws a discus. Alfonso Caramazza, Bert F. Green, Jr., and I asked college students to consider a situation in which a man ties a ball to the end of a string and whirls them in a circle above his head. What would happen to the ball, we asked, if the string were to break? We gave the students a diagram showing the circular path of the ball, and we asked them to draw the path it would follow after the string broke, ignoring air resistance. We told them to show only the horizontal motion of the ball and not to try to depict the ball's downward motion.

It is a fundamental principle of Newtonian mechanics that an object moves in a straight line unless it is acted on by an external force. Therefore the correct answer to the problem is that the ball travels in a straight line along a tangent to its original circular path, beginning at the moment the string breaks. Roughly a third of the 50 students to whom we presented the problem, however, drew a curved trajectory for the ball. That is, they believed the curvilinear motion of the ball would persist for some time even after the ball was released from the constraint imposed by the string.

Experiences reported by some of our subjects soon led us to suspect that intuitive beliefs about motion play a role not only in people's thinking about hypothetical situations but also in their interactions with real objects. For example, a subject in one of our studies told us that the first time he threw a stone with a sling he broke a window because after the stone was released it did not curve as he expected it to. In order to investigate the relation between beliefs and actions systematically, my student Deborah Kohl and I devised a simple situation in which intuitive ideas about circular motion might influence how people performed a task.

We asked a group of subjects to push a small object across a table in such a way that it would pass through a 90-degree segment of a circular ring painted on the table. The object was a small "puck" with a ball bearing that allowed it to roll smoothly across the tabletop. The ring segment was bounded by two concentric circular arcs and by two line segments that would meet, if they were extended, at the common center of the arcs. We told our subjects to start with the puck near the edge of the table. They were to push it up to one of the straight lines on the boundary of the ring segment and release it there; the puck was then to roll across to the other straight edge of the ring segment without touching either one of the arcs.

Many of our subjects realized that the puck could pass through the ring segment in a straight line, and so they tried to push the puck along a straight trajectory. A fourth of the subjects, however, moved the puck in a curved path before releasing it, in the belief that it would continue to curve after release and so would follow the arc of the ring segment. Most of the people who tried the second strategy were visibly surprised when the puck failed to curve.

It was relatively easy to construct a task in which intuitive beliefs about the motion of dropped objects might influence a person's actions. My student Allyson Washburn and I asked undergraduates to walk across a room and while they were walking to drop a golf ball so that it would hit a target marked on the floor. Almost half of the subjects tried to release the ball when it was directly over the target, in the belief that it would fall straight down. One subject deliberately walked past the target before dropping the ball, thinking the ball would move backward as it fell.

The failure to understand the role a carried object's initial motion plays in determining its path when it is dropped or thrown may also lead to difficulties with other actions. Inexperienced basketball players systematically shoot the ball much too hard when they drive toward the basket, often slamming the ball against the backboard. Novice football quarterbacks often find it hard to throw a pass accurately while they are running. A recent story in the Baltimore *Sun* described the attempts of a man who could juggle while he was riding a unicycle to teach the skill to another unicyclist. The juggler told the trainee to throw the balls straight up as she moved forward on the unicycle, but the trainee questioned the appropriateness of the instruction on the grounds that the balls would fall behind her as the unicycle moved forward.

Why do people so often misjudge the path of a moving object when they solve problems or carry out actions? Several recent studies, including the ones I have described, indicate that the errors are not random but systematic. They arise from a general, coherent theory of motion that adequately guides action in many circumstances but is nonetheless at variance with the principles of Newtonian mechanics. It is therefore the misconceptions embodied in an intuitive physical theory that occasionally give rise to errors in judgment about motion. The intuitive theory bears a striking resemblance to the pre-Newtonian theory of impetus.

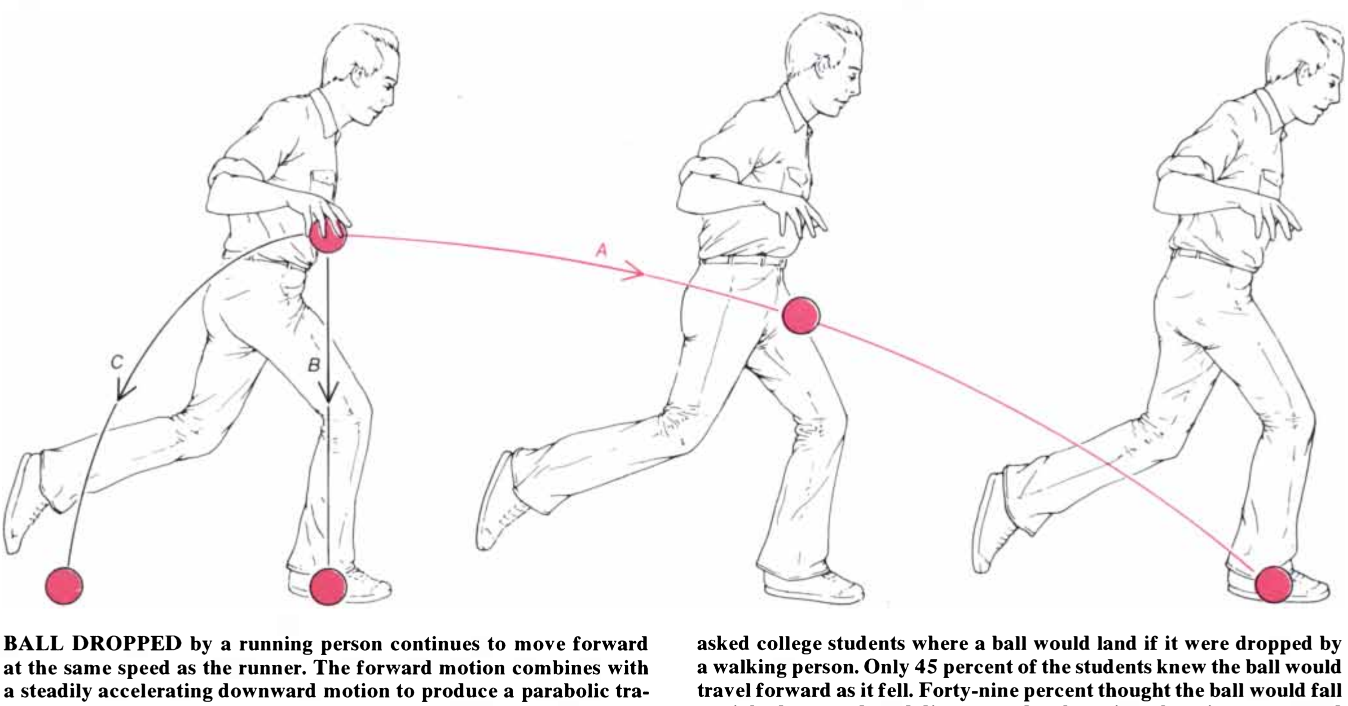
The impetus theory was a medieval correction to Aristotle's account of motion; it was discussed by the Greek scholar John Philoponus in the sixth century and was developed most fully by the French philosopher Jean Buridan and others in the 14th century. A succinct statement of the theory was given by Buridan (in translation): "When a mover sets a body in motion, he implants into it a certain impetus, that is, a certain force enabling the body to move in the direction in which the mover starts it, be it upward, downward, sideward or in a circle. It is because of this impetus that a stone moves on after the thrower has ceased moving it."

The medieval impetus theory shared with Aristotelian physics the idea that motion must have a cause. The continuous action of a force was thought to be necessary to keep an object in motion, a concept Newtonian mechanics explicitly denies. In Aristotelian physics the force responsible for the motion was necessarily external to the moving object. This requirement made it quite difficult to explain the motion of a projectile, because once a projectile leaves its launcher it is hard to find any source of external force. In the impetus theory the difficulty was circumvented by assuming that the motion of a projectile is maintained by a force internal to the object, which is acquired when the object is set in motion. The internal force was called impetus.

To account for the observation that most moving objects eventually come to a stop impetus theorists also assumed that impetus gradually dissipates. For example, according to the impetus theory, a ball rolling across a floor gradually comes to rest because the impetus acquired when it was set in motion steadily diminishes until it is entirely exhausted. Some impetus theorists held that impetus dissipates spontaneously, whereas others, including Buridan, argued that external influences such as air resistance are responsible for its loss.

The dissipation of impetus was an important consideration for the medieval analysis of projectile motion. The 11th-century Arab philosopher Avicenna held that a projectile travels in a straight line in the direction in which it is launched until the impetus from the launcher is entirely spent. The projectile then momentarily comes to rest and its natural heaviness supplies a downward impetus, causing it to fall straight to the ground. Hence for an object launched horizontally the trajectory is shaped like an inverted *L:* the object travels straight out, stops and then falls straight down.

A theory of projectile motion that gives a somewhat more accurate description of the trajectory was proposed in the 14th century by the scholastic philosopher Albert of Saxony. In Albert's account projectile motion has three phases. First the impetus of the projectile overpowers its natural heaviness, and the projectile moves along a straight line in the direction in which it was aimed. The impetus gradually dissipates, however, and at some point the projectile's weight causes it to begin to deviate downward from its original path. Thus in the second phase of its trajectory the projectile moves under



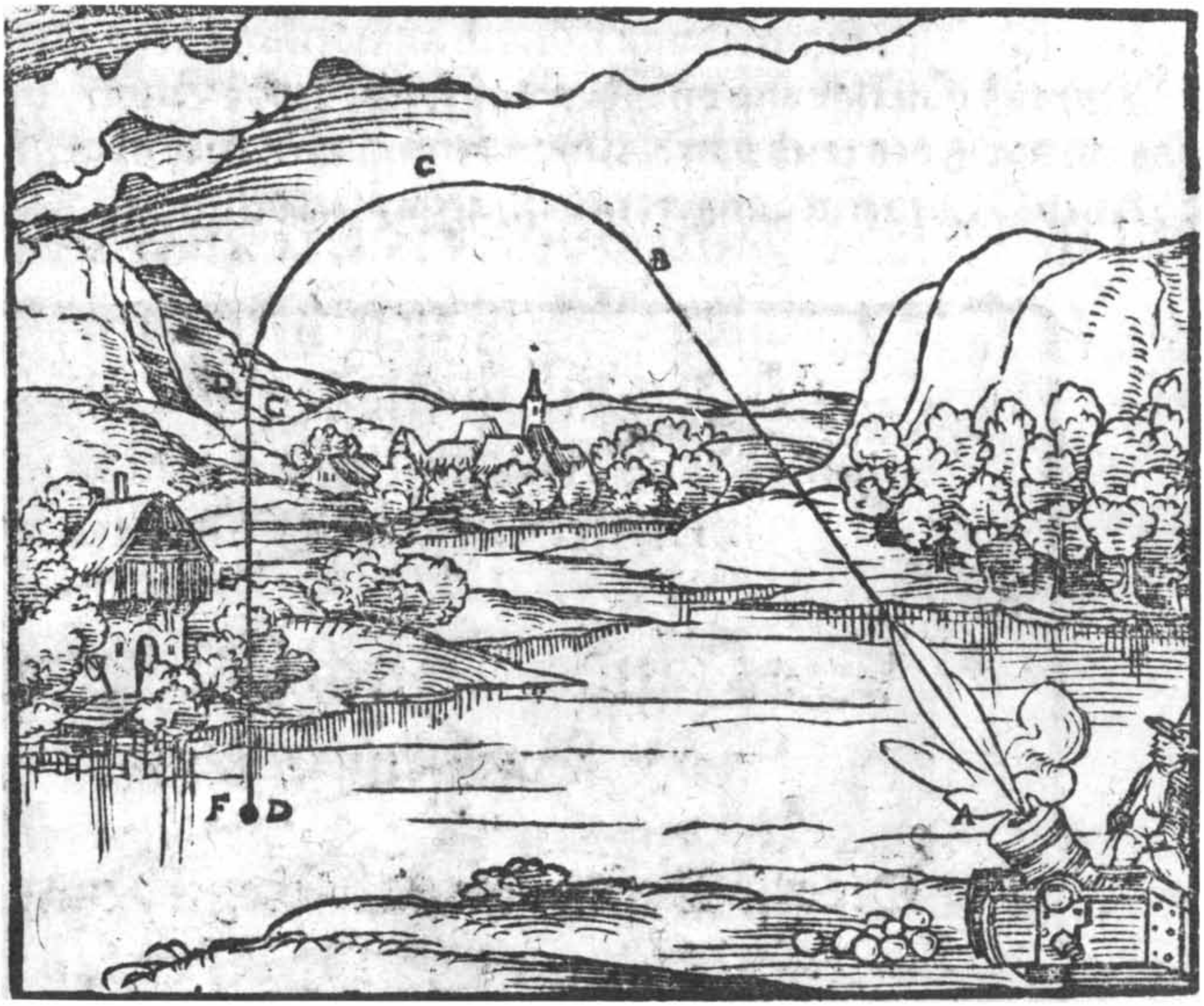
BALL DROPPED by a running person continues to move forward at the same speed as the runner. The forward motion combines with a steadily accelerating downward motion to produce a parabolic trajectory (*A*). Intuitive beliefs about the motion of objects do not always correspond with physical reality. The author and his colleagues asked college students where a ball would land if it were dropped by a walking person. Only 45 percent of the students knew the ball would travel forward as it fell. Forty-nine percent thought the ball would fall straight down and land directly under the point where it was released (*B*); 6 percent thought the ball would move backward as it fell (*C*).

the influence of both the original impetus and the downward impetus imparted by its weight. In the third phase the original impetus is entirely spent and the projectile falls straight down.

For the motion of a projectile thrown upward a similar account was given in the second century B.C. by the Greek astronomer Hipparchus. According to him, the upward impetus of the projectile initially predominates over its weight, and so the projectile moves upward; as the impetus diminishes, the upward motion slows. Eventually the upward impetus gets so weak that the weight of the projectile overcomes the impetus and the projectile begins to fall. As the upward impetus continues to dissipate, the projectile moves downward with increasing speed. In his early writings Galileo adopted Hipparchus' theory in order to discuss the acceleration of a falling object.

It is important to note that the medieval impetus theorists did not confine their concept of impetus to motion in a straight line. According to the impetus theory, an object that is moved in a circular path acquires circular impetus, which acts to maintain the circular motion. The concept of circular impetus was invoked to explain phenomena such as the continued rotation of a wheel after it is set in motion and the presumed rotation of the celestial spheres in which the stars and planets were thought to be fixed. Leonardo da Vinci applied the idea of circular impetus in order to describe the motion of an object under conditions quite similar to the ones we presented in the ball-and-string problem: "Everything movable thrown with fury through the air continues the motion of its mover; if, therefore, the latter move it in a circle and release it in the course of this motion, its movement will be curved."

The medieval impetus theory is incompatible with Newtonian mechanics in several fundamental ways. Newton's first law of motion holds that just as no force is required to keep an object at rest, so no force is required to

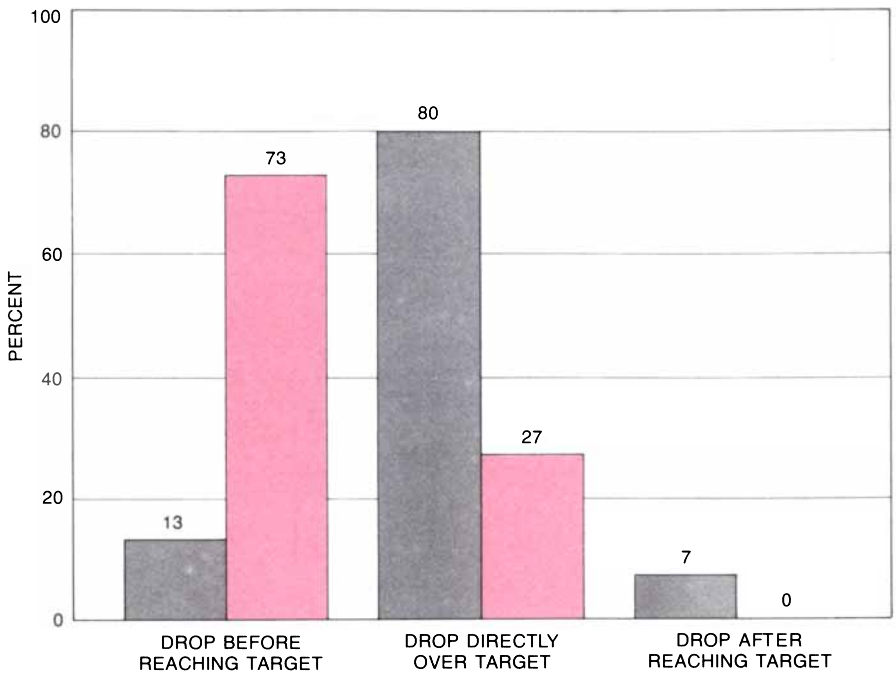


IMPETUS THEORY OF MOTION, widely endorsed in the 14th through the 16th centuries, held that when an object is set in motion, it acquires an internal force called impetus that acts to keep it moving. The theory is contrary to fundamental principles of Newtonian mechanics, which hold that no force is needed to keep an object moving in a straight line at a constant speed. Nevertheless, many people still have an intuitive understanding of the motion of objects that is quite similar to the impetus theory. This drawing, which is reproduced from the 1582 edition of *Bawkunst Oder Architectur aller fürnemsten,* by Walther Hermann Ryff, illustrates a version of the impetus theory that divides the trajectory of a projectile into three phases. In the first phase the impetus provided by the cannon overpowers the weight of the projectile, and the projectile moves in a straight line. The impetus gradually dissipates, however, and its weight causes the projectile to begin to curve downward. In the second phase the projectile moves under the influence of the impetus from the cannon and the downward impetus imparted by its weight. In the third phase the impetus from the cannon is exhausted and the projectile falls straight down.

keep an object moving at a constant velocity (that is, at a constant speed in a straight line). In fact, in Newtonian mechanics a state of rest and a state of uniform velocity are regarded as being equivalent. No physical distinction can be drawn between them, and so any object that is not changing its speed or its direction can be described as being at rest or in uniform motion, depending on the choice of a frame of reference. For example, it is equally valid to describe a passenger in a nonaccelerating airplane as being at rest in the reference frame of the airplane or as being in motion in the reference frame of the earth. Coffee poured inside the airplane falls into a cup just as it does when it is poured at rest with respect to the ground; it does not splash back toward the rear of the airplane, as it would if the impetus theory were correct.

From the perspective of the impetus theory a fundamental distinction must be made between a state of rest and a state of motion: moving objects can have impetus but objects at rest cannot. As I have pointed out, in Newtonian mechanics no such distinction can be drawn. Moving objects do have momentum, however, and so in some superficial respects the concept of impetus resembles the concept of momentum. There are two crucial differences: the first is that impetus was considered to be a cause of motion, whereas momentum is merely a quantity employed to describe motion. The second is that in the impetus theory an object is viewed as having impetus in an absolute sense, whereas in Newtonian mechanics momentum, like velocity, is defined relative to a frame of reference.

It is also incorrect to suppose a rolling object stops or a projectile falls because of the loss of impetus. In the Newtonian framework moving objects come to a stop or begin to fall because external forces act to change the speed or direction of their motion. A ball rolling across a floor is slowed by friction, a force that acts in a direction opposite to the direction of the ball's motion. A projectile fired horizontally is accelerated downward by the constant force of the earth's gravity. The projectile's horizontal motion is independent of its vertical motion, although both are slowed by the resistance of the air. In a vacuum the horizontal velocity would remain constant from the moment the projectile was launched until the moment it hit the ground. For a projectile aimed above the horizontal the continuous action of gravity steadily reduces the initially positive upward velocity to zero and then reduces it to progressively larger negative values; in other words, after the projectile reaches its peak it is accelerated downward.



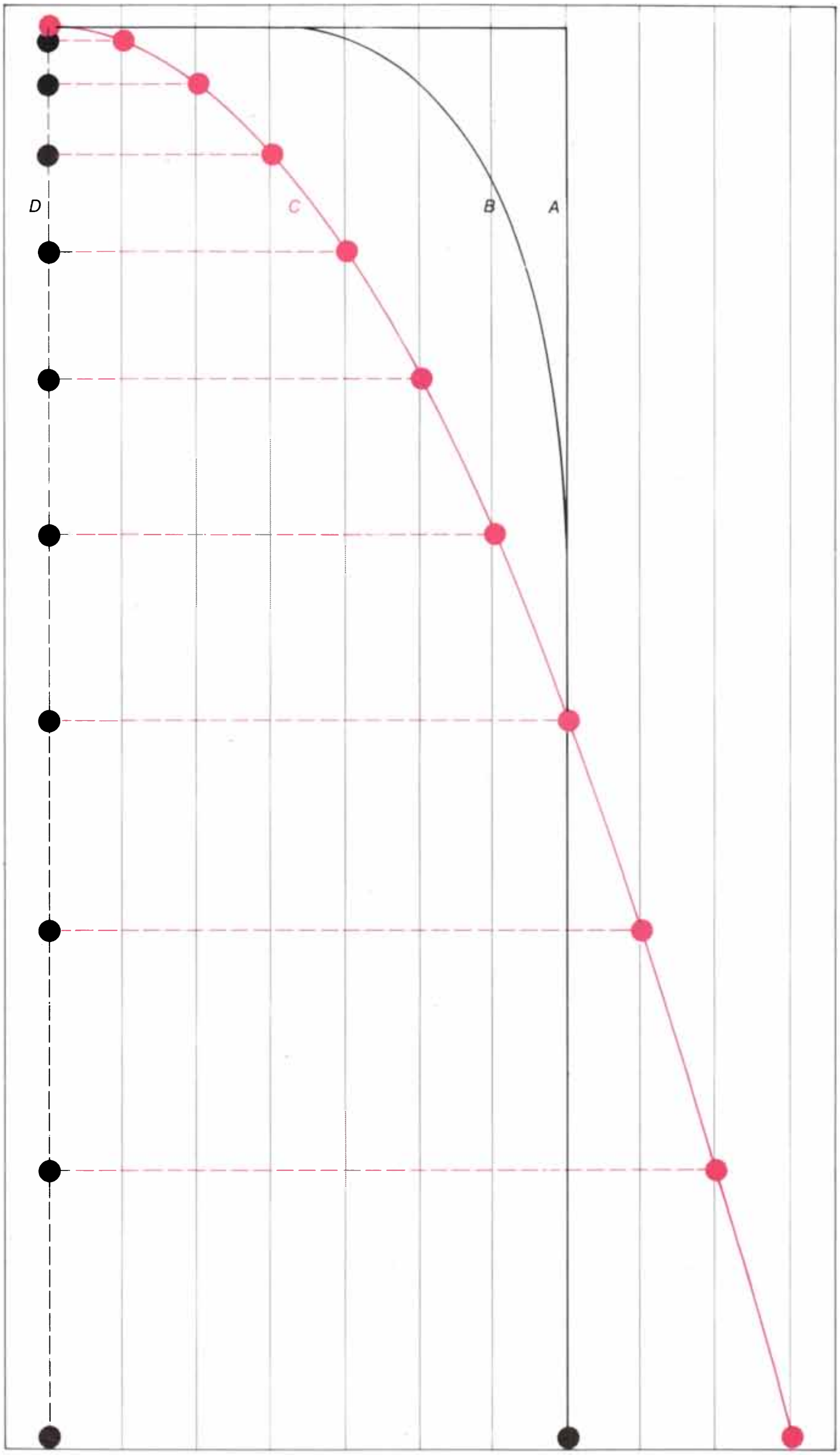
INTUITIVE BELIEFS about motion can affect the ways in which people interact with moving objects. College students were asked to walk across a room and while they were walking to drop a golf ball onto a target marked on the floor. The results of the test, summarized by the bar graphs, show that training in physics affects the strategies people adopt for hitting the target. Subjects who had taken at least one physics course (*bar graphs in color*) were likelier than those with no physics training (*bar graphs in black*) to release the ball before reaching the target. More than a fourth of the subjects trained in physics, however, incorrectly thought the ball should be dropped when it was directly over the target. The finding shows that students may often retain the intuitive impetus theory as a framework for interpreting course material on Newtonian mechanics and therefore distort the new material to fit the intuitive theory.

The effect of gravity on a projectile's vertical motion is not delayed by any horizontal component of motion; the projectile begins to deviate downward from the direction in which it was aimed the instant it is fired. Thus the impetus theorists were also incorrect in maintaining that a projectile travels along a straight line for some time after it is launched.

Finally, Newtonian mechanics and the impetus theory differ in their treatment of circular motion. In the impetus theory circular motion is viewed as being not fundamentally different from motion in a straight line. Both forms of motion are generated by imparting the appropriate impetus to an object. In Newtonian mechanics there is clear distinction between motion in a straight line and motion in a circle. An object moves uniformly in a straight line when no force is acting on it. For an object to move in a circle, however, it must be acted on constantly by an outside force that tends to deflect the motion from linearity. A ball whirled in a circle at the end of a string is acted on by the tension in the string, which continuously pulls on the ball and so keeps it from moving off at a constant speed in a straight line tangent to the circle.

Each major aspect of the impetus theory I have discussed has its counterpart among the beliefs about motion that are widely held today. Indeed, the ideas about motion held by most people with no formal training in physics, and by many who have completed at least one physics course, are much closer to the account given by the impetus theory than they are to Newtonian mechanics. These conclusions are supported by our studies at Johns Hopkins and by experiments carried out by John J. Clement of the University of Massachusetts at Amherst, by Audrey B. Champagne and Leopold E. Klopfer of the University of Pittsburgh and by Laurence Viennot of the University of Paris.

In one study, carried out by Kohl, Washburn and me, high school and college students tried to solve several problems about motion and then explained their answers. In many instances the students stated the impetus theory quite explicitly. For example, one student, who used the term momentum in explaining his answers, was asked to define the term. He responded: "It's something that carries [an object] along after a force on it has stopped. Let's call it the force of motion. It's something that keeps a body moving." Another student described the momentum of a ball: "[It is] a force that has been exerted and put into the ball so this ball



BELIEFS ABOUT PROJECTILE MOTION held by experimental subjects resemble those of medieval impetus theorists. Subjects were asked to draw the trajectory of a projectile launched horizontally, ignoring the resistance of the air. Five percent indicated that the projectile moves straight out and then straight down (*A*). The 11th-century Arab philosopher Avicenna held a similar view. Thirty-five percent of the subjects believed the projectile first travels in a straight horizontal line, then begins to curve downward and finally falls straight down (*B*). The description agrees with the three-phase theory of projectile motion proposed by the 14th-century scholastic philosopher Albert of Saxony. Twenty-eight percent of the subjects gave the correct trajectory, which is parabolic (*C*); most of the remaining 32 percent correctly indicated that the projectile begins to fall immediately but incorrectly thought it eventually falls straight down. The parabolic path represents the combination of a constant-velocity horizontal motion and an accelerating vertical motion. The distance the projectile falls during each interval of time is the same as that of an object dropped from rest (*D*). Thus a bullet fired horizontally and a bullet that is dropped simultaneously from the same height hit ground at the same time.

now that it's traveling has a certain amount of force. The moving object has the force of momentum and since there's no force to oppose that force, it will continue on until it is opposed by something." In the two statements the belief that motion is maintained by an internal force is quite clear.

Also expressed by our subjects was the idea that moving objects slow down and stop because of the dissipation of impetus. One subject, asked to explain why a ball rolling along the floor would eventually come to a stop, gave the following explanation: "I understand that friction and air resistance adversely affect the speed of the ball, but not how. Whether they sort of absorb some of the force that's in the ball—I'm not sure. In other words, for the ball to plow through the air resistance or the friction, if it has to sort of expend force and therefore lose it, I'm not sure." The speaker clearly, although tentatively, assumes that friction and air resistance dissipate the impetus of the ball.

The intuitive idea that an object released from circular motion follows a curved path for some time thereafter appears to stem from a belief that such an object has circular impetus. A subject who drew a curved trajectory for our problem involving a ball whirled at the end of a string explained that the ball would curve "because of the directional momentum. You've got a force going around, and [after the string breaks, the ball] will follow the curve you've set it on until the ball runs out of the force within it that you've created by swinging." In other words, the whirled ball acquires circular impetus, which causes it to follow a curved path for some time after the string breaks. The circular impetus dissipates, however, and so the trajectory of the ball gradually straightens out.

Our subjects also recapitulated the medieval impetus theory in their beliefs about projectile motion. When subjects were asked to explain why projectiles rise to a peak and then fall, they frequently echoed the account initially proposed by Hipparchus. For example, one subject explained the vertical motion of a cannonball in the following way: "As the cannonball comes up, the force from the cannon is dissipating and the force of gravity is taking over. So it's slowing down. At the peak the force of the cannon and the force of gravity are fairly equal. That's the changeover point, at which point the force of the cannon is less as gravity becomes the more predominant force. As it makes the arc and begins to come down, gravity is overcoming the force from the cannon." Clement found that students give similar explanations for the motion of a coin tossed upward.

Some subjects who gave this kind of

explanation stated that a projectile accelerates as it falls because of the continuing dissipation of the original upward impetus. Perhaps it bears repeating that according to Newtonian mechanics, there is no upward impetus and no upward force; there is only the upward component of the projectile's velocity, which is steadily reduced by the action of gravity. The peak of the projectile's arc does not represent a point at which upward and downward forces are equal, as it does in the impetus theory.

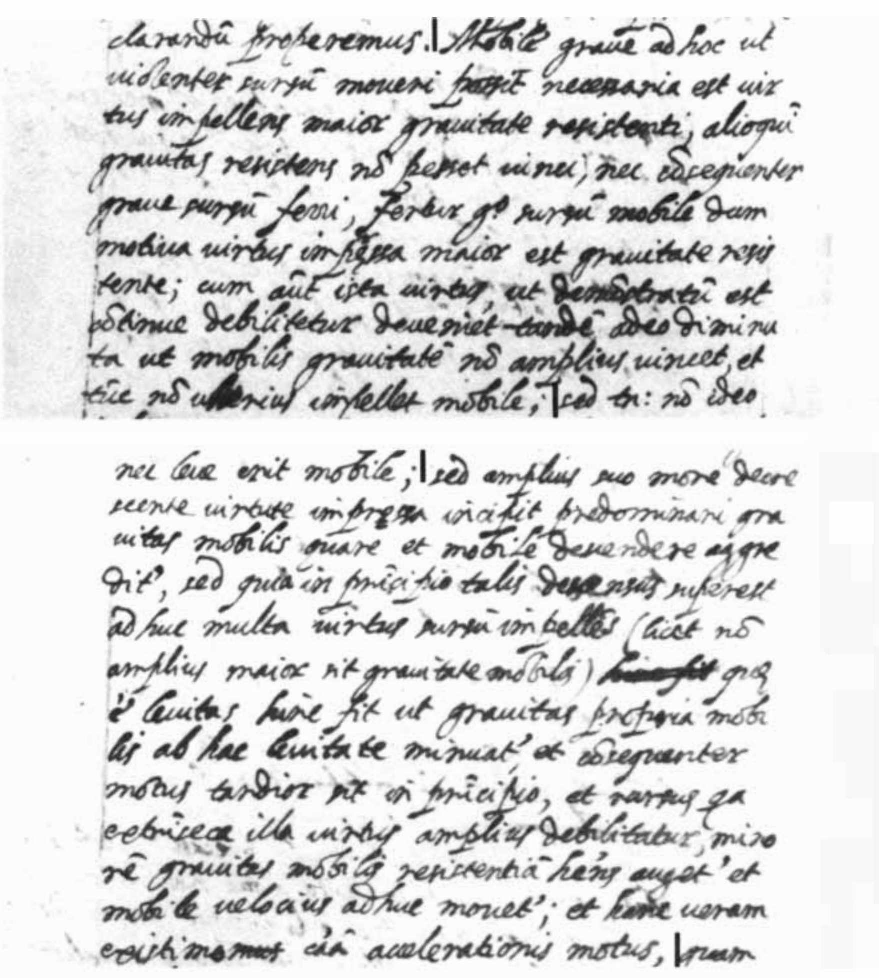
In one of our problems on projectile motion we asked the subjects to suppose a metal ball is pushed along the top of a cliff at high speed and then goes over the edge. Their task was to draw the path of the ball as it falls, ignoring air resistance. The correct answer is that the ball's constant horizontal velocity and its steadily increasing vertical velocity combine to generate a parabolic arc.

More than a third of our subjects who were given the cliff problem unknowingly took the position of Albert of Saxony: the ball, they maintained, travels in a straight horizontal line for some distance beyond the edge of the cliff. It then begins to curve downward and eventually falls straight down. One subject explained: "When [the ball] leaves the cliff, the horizontal force is greater than the downward-motion force. When the horizontal force becomes less, the ball would start falling." Five percent of our subjects echoed Avicenna's theory: the ball, they said, travels straight out and then straight down. One subject who gave this answer defended his reasoning by saying that "gravity isn't going to affect it until it stops moving."

The belief that a carried object falls straight to the ground when it is dropped also derives from the intuitive impetus theory. Subjects who hold this belief explain that an object acquires impetus when it is pushed or thrown but not when it is passively carried. Therefore, according to the impetus theory, when a carried object is dropped, there is no force to cause it to move forward, and so it falls straight down. The intuitive belief is maintained even for situations where the carrier is moving at a high speed. In one experiment we asked students to draw the trajectory of a bomb dropped from a moving airplane, ignoring the effects of air resistance. More than a third of the subjects indicated that the bomb would fall in a straight vertical path.

The striking similarity between the views of the medieval philosophers and those of our subjects suggests the impetus theory is a natural outcome of experience with terrestrial motion. It is easy to see how experience could lead to the basic idea that objects are kept in motion by an internal force that gradually dissipates: in most circumstances an object that is set in motion keeps moving for some time, but it eventually stops. What is difficult to understand is why people develop incorrect beliefs about the trajectories of moving objects, beliefs that apparently conflict with everyday experience. Why, for example, do people come to believe carried objects fall straight down when they are dropped?

A possible answer is that under some conditions the motion of objects is systematically misperceived. Washburn, Linda Felch and I have proposed that objects dropped from a moving carrier are often perceived as falling straight down. The misperception may be the source of the corresponding belief that they do so. Studies in the perception of motion have shown that when an object is viewed against a moving frame of reference, a visual illusion can arise. The motion of the object relative to the moving frame of reference can be misper-

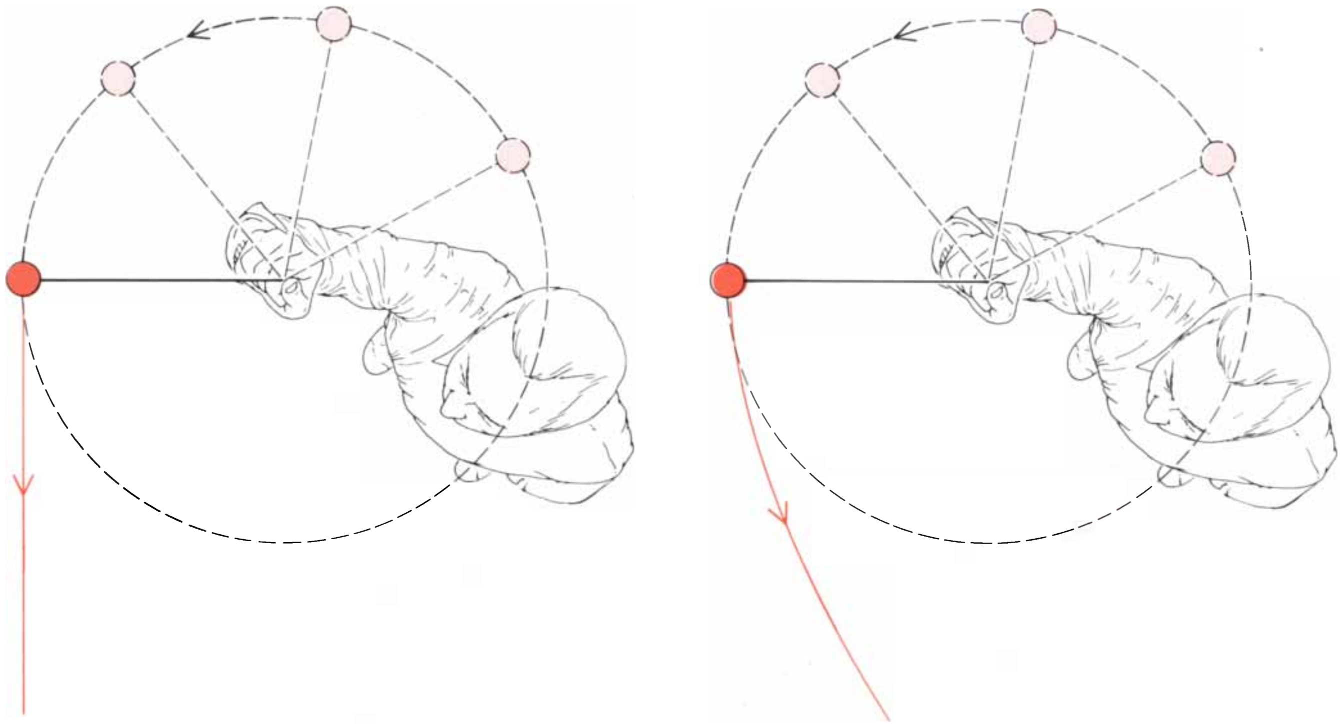


For a heavy body to be able to be moved upward by force, an impelling force greater than the resisting weight is required; otherwise the resisting weight could not be overcome and consequently the body could not move upward. That is, the body moves upward, provided the impressed motive force is greater than the resisting weight. But since that force, as has been shown, is continuously weakened, it will finally become so diminished that it will no longer overcome the weight of the body and will not impel the body beyond that point. . . .

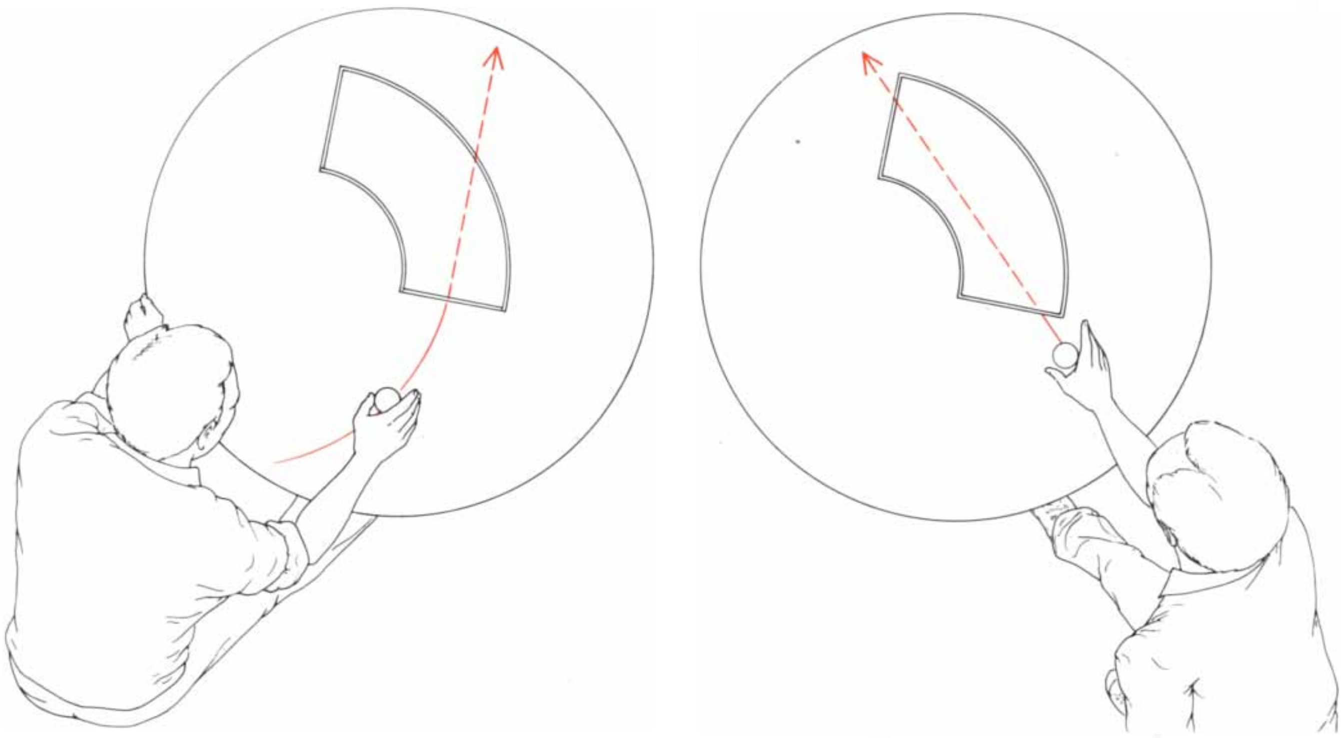
But beyond that, as the impressed force characteristically continues to decrease, the weight of the body begins to be predominant and consequently the body begins to fall. Yet there still remains, at the beginning of the descent, a considerable force that impels the body upward, which constitutes lightness, although the force is no longer greater than the heaviness of the body. For this reason the essential weight of the body is diminished by the lightness and consequently the motion is slower at the beginning. Furthermore, since that external force continues to be weakened, the weight of the body, being offset by diminishing resistance, is increased and the body moves faster and faster.

This is what I consider to be the true cause of the acceleration of motion.

MANUSCRIPT OF GALILEO'S *DE MOTU* ("On Motion"), an essay written in about 1590, shows that he endorsed the impetus theory in his early writings. In the excerpt Galileo applies the concept of impetus to explain the motion of a projectile launched upward. The author and his colleagues at Johns Hopkins University and John J. Clement of the University of Massachusetts at Amherst have found that experimental subjects who are asked to explain the rise and fall of a projectile frequently give quite similar accounts. The manuscript is in the Biblioteca Nazionale Centrale in Florence; the English translation was done by the late I. E. Drabkin.



CIRCULAR IMPETUS, according to the impetus theory, is imparted to an object when the object is moved in a circle. Hence the impetus theory incorrectly predicts that a ball whirled at the end of a string will continue to follow a curved path if the string breaks. Actually the ball will move in a straight line along a tangent to its original circular path, beginning at the instant the string breaks. Fifty-one percent of the college students who were asked to draw the path of the ball after the string breaks sketched the path correctly (*left*). Thirty percent, however, thought circular motion would persist and the path would be curved (*right*); 19 percent gave other incorrect responses.



SOME PEOPLE INTERACT WITH MOVING OBJECTS as if the objects could be given circular impetus. A 90-degree segment of a ring was painted on a table, and experimental subjects were given a small "puck" with a ball bearing that would enable it to roll smoothly across the table. The subjects were asked to push the puck up to one edge of the ring segment and release it; the task was to make the puck cross to the other side of the segment without touching the curved sides. Twenty-five percent of the subjects tried the strategy at the left: they moved the puck in an arc, apparently in the mistaken belief that the object would continue to travel in a curved path as it moved through the ring segment. The broken line, which represents the path of the puck after it was released, shows that this strategy invariably failed. Sixty-seven percent of the subjects applied the correct strategy that is shown at the right: they aimed the puck so as to take advantage of its straight-line trajectory after it was released. The remaining eight percent of the subjects tried other unsuccessful strategies.

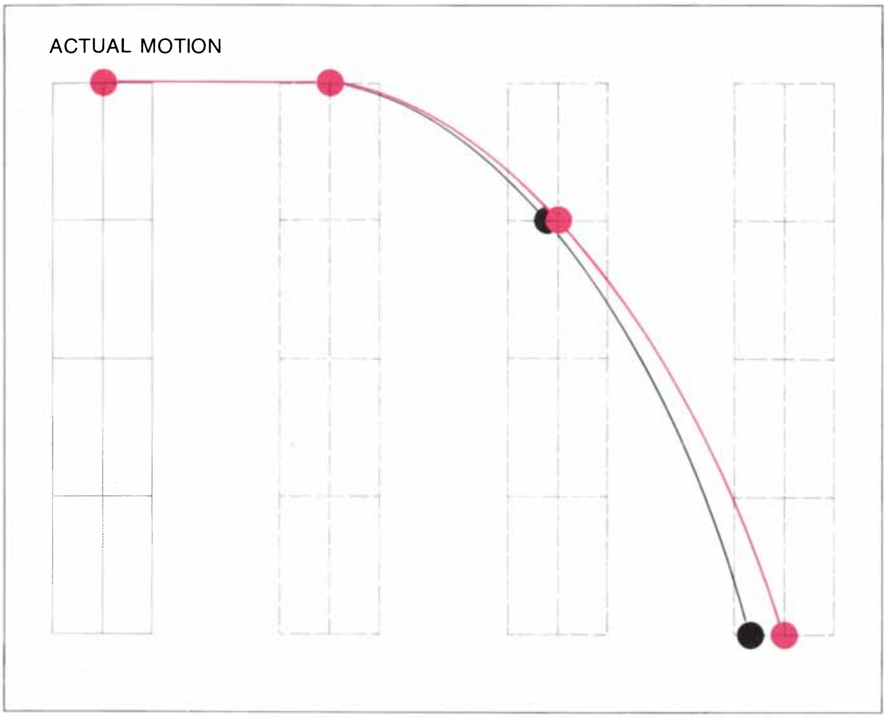
ceived as absolute motion (that is, as motion relative to the stationary environment). For example, if a dot in the interior of a rectangle remains motionless as the rectangle moves to the right, the dot can be perceived as moving to the left [see "The Perception of Motion," by Hans Wallach; SCIENTIFIC AMERICAN, July, 1959].

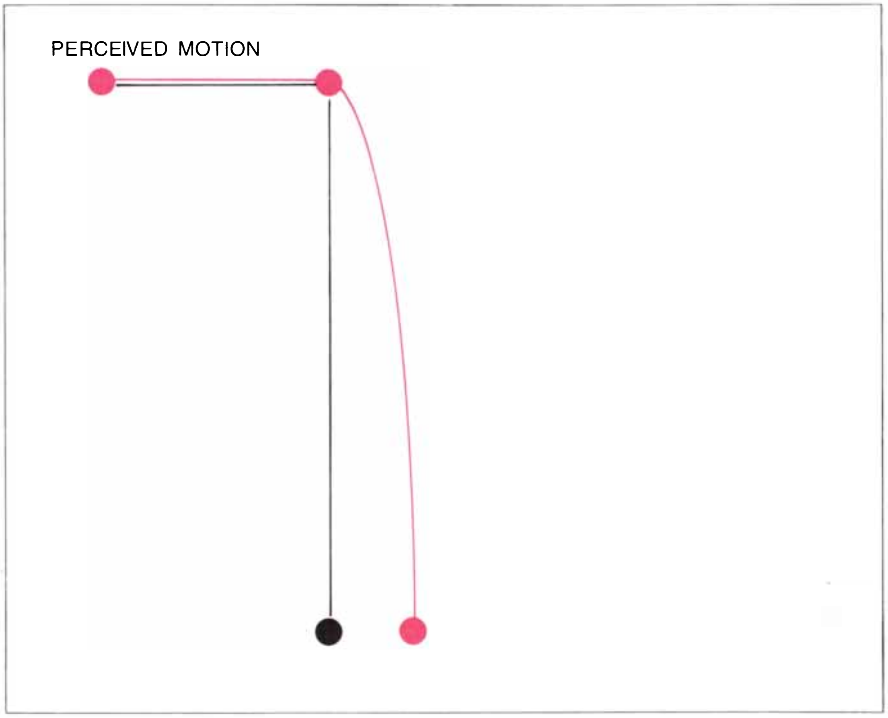
Similarly, when people observe carried objects that are dropped, the moving carrier usually remains in view as the dropped object falls; hence the carrier can act as a frame of reference. In situations where the effects of air resistance are negligible, the dropped object falls straight down with respect to the carrier. Therefore the illusion could develop in which the object appears to fall in a straight vertical trajectory with respect to the ground.

In order to test this hypothesis we asked subjects to watch computer-generated displays resembling situations in which carried objects are dropped. A rectangular box, representing the carrier, moved across a video screen at a moderate, constant speed from left to right. A round dot, representing the carried object, remained at the top center of the moving box until the box had traveled about a third of the distance across the screen. The dot then traced one of several paths that were designed to simulate the trajectory of an object dropped in the presence or absence of air resistance. The box kept moving at its initial speed while the dot fell. For each display the subject first watched the motion and then tried to draw the path that had been taken by the dot.

Our results indicated that subjects experienced a strong illusion when they viewed the displays. When the dot traced the path of an object falling in a vacuum, our subjects generally drew a path that was much steeper than the one they had just observed. That is, they perceived the path of the dot as being nearly vertical and curved only slightly in the direction of the motion of the box, even though the dot had actually moved much farther to the right. When the dot fell like an object affected slightly by air resistance, it was generally perceived as falling straight down, although again it actually moved considerably farther to the right during its fall. The illusion persisted only when the moving box was shown; when the subjects observed the same motion of the dot in the absence of the moving frame of reference, the illusion did not arise and the drawings were accurate.

In a second experiment our subjects watched a videotape of a walking person dropping a ball; again their task was to draw the path of the ball. We warned them that the videotaped event might have been faked through trick photogra-





ILLUSIONS IN THE PERCEPTION OF MOTION may explain some of the widespread misconceptions people have about moving objects. The author and two of his students, Allyson Washburn and Linda Felch, tested the relevance of this hypothesis for explaining the incorrect belief that an object dropped by a moving carrier falls straight down. A rectangular box, representing the carrier, moved from left to right across a video screen. A dot, representing the carried object, remained fixed at the top center of the box for about a third of the distance across the screen and then began to fall. The falling dot moved like an object dropped in a vacuum (*colored curve in upper illustration*) or like an object affected slightly by the resistance of the air (*black curve in upper illustration*). After watching the motion of the dot and the rectangle experimental subjects tried to draw the path the dot had just traced with respect to the screen. When the dot fell as if it were in a vacuum, it was generally perceived as moving only slightly forward in the course of its fall (*colored curve in lower illustration*). When the dot moved as if it were influenced by the resistance of the air, the average perceived path was a straight vertical line (*black curve in lower illustration*). There was no such illusion when subjects viewed the same motions of the dot in the absence of the moving rectangular background.

phy, although actually it was not. Consequently, we told them, they were to draw exactly what they observed and not what they thought should have happened. Of the 18 subjects who viewed the videotape only four indicated that the ball moved forward as it fell. Ten subjects said the ball fell straight down and four thought it moved backward.

Such results do not prove that misperception is the cause of mistaken beliefs about the motion of dropped objects, but they do suggest strongly that the hypothesis is a reasonable one. My colleagues and I are exploring the possibility that illusions are implicated in the development of other misconceptions about motion.

How can such misconceptions be dispelled? The most obvious answer is through normal instruction in Newtonian mechanics, but studies by several investigators suggest that the intuitive ideas are difficult to modify. Although some students who take physics courses achieve a firm grasp of Newtonian mechanics, many emerge with their intuitive impetus theories largely intact. Kohl, Washburn and I tested the knowledge of motion among high school students before and after they had finished a physics course and found that the course had left some misconceptions unaffected. The instruction was particularly unsuccessful in altering the core assumption that impetus acquired when an object is set in motion serves to maintain the motion: 80 percent of the students retained this belief even after finishing the course. (Ninety-three percent held the belief prior to instruction.)

It appears that students often rely on the intuitive impetus theory as a framework for interpreting new course material. As a result the material can be misinterpreted and distorted to fit the intuitive preconceptions. Several educators are seeking ways of designing physics instruction to avoid such pitfalls. James Minstrell, a high school teacher in Mercer Island, Wash., has reported some success with a method whereby students are led to articulate their intuitive beliefs. The differences between such beliefs and the principles of Newtonian mechanics can then be sorted out. Work by Andrea A. di Sessa of the Massachusetts Institute of Technology suggests that experience with computer games, in which objects act as if they moved in a frictionless Newtonian world, may also be helpful.

Hence the study of intuitive theories and the processes by which they are acquired and modified holds promise for the development of improved educational methods. Such study may also help to illuminate the factors that influence the historical development of formal scientific theories.

THE AMATEUR SCIENTIST

How lenses can be made out of ice, and what happens when coffee is brewing in the ibrik

by Jearl Walker

A lens made out of ice seems at first thought to be improbable. It is quite practical, however, and offers a number of avenues for experiment. The notion figured prominently in one of Jules Verne's science-fiction tales, *The Desert of Ice*. Now it has been brought to reality by Matthew G. Wheeler of McBride, B.C.

In Verne's story a group of explorers survive a catastrophe in the Arctic by means of an ice lens. The ship's captain had set out to reach the North Pole, but a mutiny had left him and a few supporters with a broken and sinking ship and few supplies. Their only chance to survive was to reach another ship that had been abandoned when it became locked in the ice. The group gathered wood from their broken ship, packed up their meager food supplies and the flint and steel for making fire and set out.

Because their food was running out, they stalked and killed a polar bear, but as they were doing so their fire went out and the flint and steel were lost. Although they had enough food for the remainder of the trip, they were threatened with freezing to death if they could not restart the fire.

They were saved by optics. Noting the bright sunshine, a physician in the group recalled how a fire can be started with a lens. Lacking a glass lens, he decided to make a lens of ice. He chose a clear section of ice and hacked out a piece about a foot in diameter. Shaping it roughly with a hatchet, he whittled and smoothed it with his knife and then polished the surface with his bare hands "until he had obtained as transparent a lens as if it had been made of magnificent crystal." The ice lens was positioned above some tinder, focusing the rays of the sun. Within a few seconds the tinder had caught fire. "The stove was soon roaring, and it was not many minutes before the savory odor of broiled bear-steak" had roused the men.

What the physician had made was a convex lens. Since it was a large one, it intercepted a goodly amount of sunlight. The concentration of the light on the tinder placed at the focal point of the lens quickly delivered enough energy to kindle a fire.

I told this story on one of my regular contributions to Quirks and Quarks, a radio show about science on the Canadian Broadcasting Corporation. I also described how one can fashion a convex lens of ice in a simpler way with curved watch glasses. (A watch glass is the kind of glass that covers the face of a watch. Glasses of similar shape but larger size can be found among the standard supplies of a chemistry laboratory.) Water is poured into two watch glasses and allowed to freeze. Then the flat sides of the two pieces of ice are placed together to form the lens.

Wheeler was a listener who decided to improve on my idea. He made a lens from the ice in a cattle trough. He started with a section that was about an inch thick and free of bubbles. To shape the piece he put it in the lid of a milk pasteurizer, which is curved like a lid for a saucepan. Then he floated the lid (with the piece of ice) in a sink of warm water. The heat melted the bottom of the ice until it fitted the inside of the lid. He occasionally rotated the ice to increase the symmetry of its bottom surface. Next he turned the piece over to shape the other side.

Wheeler now had a convex lens of ice. To test its focusing power (and to mimic the physician in the Verne story) he held the lens by its edge and focused the light of the sun onto a newspaper: "The paper quickly singed, smoked and burst into flame." Wheeler tells me the lens worked best when the air temperature was above freezing. At lower temperatures drops of water condensed on the lens, distorting its surface and interfering with its focusing of light.

Wheeler made a smaller lens in the concave base of a pressurized container. When he placed the lens in front of his 35-millimeter camera, it functioned as a telephoto lens with a focal length of about 80 millimeters. Wheeler tested it outdoors on a cold day with snow on the ground. The lens was too fast to photograph snow scenes, and so he stopped it down to $f9$ with a cardboard iris. The iris was mounted between a collar of black plastic tubing and a conventional rubber lens shade. The purpose of the collar was to eliminate any light that did not pass through the lens.

Initially Wheeler held the lens in place with his gloved hand. Later he found that a cardboard mount was more convenient. The lens was clamped to it with three cardboard flanges stapled to the card. On the front of the card he put a cardboard shade. Once the lens was mounted in its holder the card and the lens were held against the rubber lens shade and adjusted for position. Wheeler told me an ice lens will work well for 15 minutes and will serve even when the air temperature is 16 degrees Celsius (about 60 degrees Fahrenheit).

Wheeler has now made many lenses with a variety of molds. In some cases he begins with liquid water that he freezes. He says the clearest ice results when the freezing begins at the top. To make this kind of ice he insulates the sides and the base of his mold.

If you make an ice lens with a pair of watch glasses, experiment with placing the two pieces of ice together to complete the lens. The pieces should have good contact. Try rubbing them together or stroking them with a finger so that the contact surfaces are wet and smooth. When they freeze together, the contact should be firm.

By choosing appropriate molds you can fashion both convex and concave lenses or combinations of the two. The index of refraction for ice is less than that for glass, so that an ice lens does not bend the rays of the sun as much as a glass lens of the same shape. Therefore an ice lens (either convex or concave) has a longer focal length than a similar lens made of glass.

Middle Eastern coffee is a strong drink brewed from sugar and finely ground coffee. The sugar is needed to offset the bitter taste of the coffee, but I think it also is a factor in the brewing. A mixture of ground coffee, sugar and water is heated in an ibrik, a container that tapers slightly toward the top and has a long handle. The coffee is rapidly brought to a boil and then poured into a demitasse, which is half the size of a standard coffee cup. What is poured is a mixture of liquid, foam and grains of coffee. The foam is highly regarded in Middle Eastern countries. In fact, some people say it is an insult to serve a guest a demitasse of coffee without foam.

One must let the coffee remain in the demitasse for several minutes before it is tasted. The grains of coffee settle slowly to the bottom. If the brew is sampled too soon, the liquid is unpleasantly gritty. After the grains have settled the liquid is sipped slowly until nothing

Matthew G. Wheeler employing one of his ice lenses to ignite paper

A scene in British Columbia photographed by Wheeler with an ice lens mounted on a camera remains but the bottom layer of coffee grains, which is about the consistency of a sludge.

I have spent hours meditating on the brewing and sedimentation of Middle Eastern coffee as I prepare demitasse after demitasse of the drink. It seems to be about the only coffee drink in which grains of the coffee remain in the liquid after the brewing process. It is also noteworthy in several other respects. Why must the coffee be finely ground? (When you buy coffee for the drink, you should ask for an extra-fine grind.) Why is sugar heated with the coffee rather than added later as it usually is with other coffee drinks? Why does the liquid suddenly foam up the sides of the ibrik as the coffee is being brewed? How long must one let the coffee stand in a demitasse to ensure an ungritty sip?

I buy finely ground coffee roasted by the Viennese or the French method. Both coffees are darker than the others commonly sold in the U.S., meaning that they have been roasted longer and have lost more of the volatile substances in them. Nevertheless, I doubt that the type of coffee alters anything but the taste of the drink.

I prepare my coffee in a tin ibrik large enough to fill two demitasses, each of which holds about two fluid ounces (roughly 60 milliliters). Usually, however, I make only a single demitasse at a time, otherwise the foam forms so fast that it cascades onto the stove. Hence I add a bit less water to the ibrik than is needed to fill one demitasse. I add two teaspoons of coffee grounds and from one teaspoon to four teaspoons of sugar to the water. (One teaspoon is equivalent to five milliliters.)

Some people mix the ingredients before brewing the drink, but I usually do not. I put the ibrik on the heating coil of my electric stove with the control set on high. Soon the mixture begins to rumble and click. After several minutes the noise suddenly stops and the foaming starts. The foam rapidly builds upward to the top of the ibrik. The transition is so fast that I sometimes fail to remove the ibrik from the heating coil before the foam overflows. If I pour the coffee slowly into the demitasse, some of the foam remains in the cup, but most of it collapses.

In order to look inside the mixture while it was brewing I replaced my ibrik with a 180-milliliter beaker made of Pyrex. Again I added the proper amount of water and coffee for one demitasse. The coffee grounds floated and very slowly became wet. Apparently they are held together with enough electric force to keep them in a clump. Water seeps between the grains only slowly. Although the grains are denser than water and so should sink, the clumps contain enough trapped air to make them float.

I put two teaspoons of sugar onto the coffee. Most of the sugar sank, sometimes tipping over a clump of coffee as it did so. When I placed the beaker on the heating coil, the sugar and coffee layers were well separated by a layer of relatively clear water.

Even if I stir the mixture before heating it, the same three layers appear. In cold water the sugar dissolves slowly and so settles to the bottom after being stirred because it is denser than water. Even with stirring, the coffee grains still carry enough adsorbed air to make most of them float to the surface. The only difference when the mixture is stirred is that some of the grains end up in the water layer.

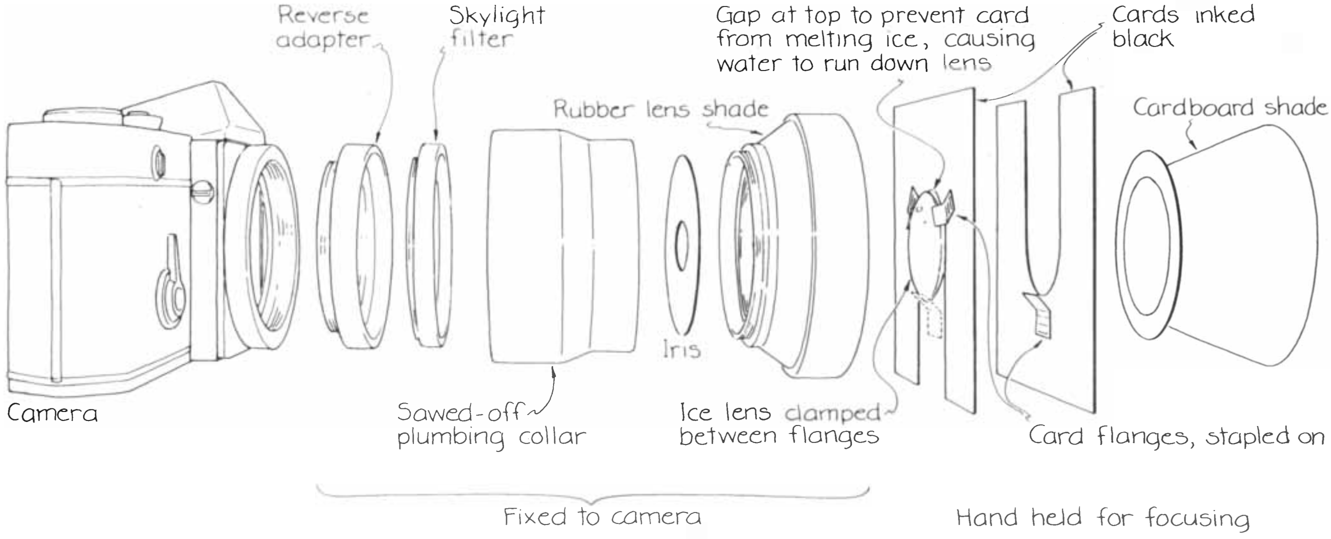
Soon after the heat was turned on bubbles of air appeared at several sites along the bottom of the sugar layer and began oscillating rapidly. The air in the bubbles had been adsorbed onto the sugar grains or had been dissolved in the water between the grains. Once bubble nucleation starts at a given place more air is added to the site from the surrounding water and sugar. The bubbles oscillate because when they expand upward into cooler liquid (concentrated sugar water in this case), they suddenly collapse. Soon afterward, however, they are reheated and reexpand. The rapid and chaotic oscillation of the air bubbles generates the clicks and rumbles characteristic of the early stage in heating a fluid.

Bubbles of air and water vapor eventually break free and stream upward through the sugar layer. When they penetrate the top of the layer and enter the clear water, their motion is arrested. The water is cooler than the sugar layer and makes the bubbles collapse.

By shining a flashlight into the beaker I can spot threads of sugar water projected upward from the collapse of each bubble. The threads are visible because their index of refraction differs from that of the water. The threads twist and turn before they fall back to the sugar layer. Sometimes small droplets of sugar (or concentrated sugar water) are left in the wake of the threads.

Soon the entire sugar layer becomes turbulent. Whereas until then bubbles conveyed heat upward to the layer of water, convection streams now seem to dominate. The variation in the index of refraction makes some of these streams visible. With further heating the turbulent sugar layer moves upward into the water layer. Some of the turbulence or one of the bubbles from the top of the sugar layer occasionally reaches the bottom of the coffee, but otherwise the coffee seems to be unaffected by what is taking place below it.

As the turbulent layer of sugar nears the coffee the grains begin to glisten with



Wheeler's arrangement for mounting an ice lens on a camera tiny air bubbles, none larger than a millimeter and most much smaller. Just as the sugar layer reaches the coffee the grains are stirred around and yield a great many of the tiny bubbles. They form the foam, which builds up rapidly. Within a few seconds the foam is three times as high as the initial mixture.

As soon as the grains are stirred around the rumbles and clicks stop. The noise originated in the formation of large bubbles along the bottom of the container. Once the grains reach the region they provide plenty of nucleating sites for smaller bubbles. The large, noisy bubbles disappear.

I employ a long pair of tongs to grasp the hot beaker. If I hold the beaker above the heating coil, I can control the rate of production of the foam: the rate slows as I move the beaker away from the coil. The bubbles are small at first, but as they are pushed up the beaker by newly made bubbles they merge and give rise to larger bubbles. The larger bubbles are too fragile to survive when the coffee is poured. The foam some people like to have on their Middle Eastern coffee consists of the tiny bubbles that were adhering to grains of the coffee when the foaming started. When the coffee is poured into a demitasse, some of the tiny bubbles break free of the grains and form a raft that may hold its shape for quite a while.

I do not always get such a raft on my coffee. When I do, it can survive for hours if the demitasse is left undisturbed. One reason for the stability of the bubbles is that the sugar in the coffee increases the viscosity of the fluid. When a bubble raft has formed, the fluid in the walls between the pockets of gas drains because of gravity. The walls get thinner until the bubbles burst or merge. The more viscous the fluid is, the slower it drains. The fluid in the walls of the bubbles on my coffee drains very slowly.

To change the conditions I made the coffee without the sugar. Again the coffee grains floated in clumps on top of the water. When the beaker was placed on the coil, the water began to heat rapidly and with much noise. Air bubbles appeared and oscillated. Then bubbles of water vapor formed and escaped upward to beat against the bottom of the coffee layer.

Soon streams of these bubbles broke the clumps of coffee. The grains quickly dispersed throughout the volume of the fluid. Although the clumps lost some of their air and thus their buoyancy, the grains still bore adsorbed air. They were stirred down from the top surface only because of the strong convection currents in the beaker. Small bubbles formed and then the foaming began. All these things happened faster than they did when I had sugar in the mixture.

I tried several other preparations. In one I heated only sugar (two teaspoons) and water. As before, the sugar sank to the bottom and was heated first. Viscous bubbles and turbulent convection appeared again. When the mixture was at full rolling boil, bubbles appeared on the top, but with this preparation there was no rapid foaming. The bubbles were relatively large (several millimeters across) and did not last long.

After a few minutes the fluid was transparent, indicating that the sugar had dissolved fully and was uniformly mixed. I allowed the mixture to cool. Even at room temperature all the sugar remained dissolved. Clearly the amount of sugar normally added to the drink does not exceed the saturation limit of sugar in hot water.

Next I boiled the usual amount of water in the beaker but omitted the sugar. After the rolling boil began I added a single teaspoon of coffee. Almost immediately the grains were swirled throughout the volume of water and tiny bubbles developed on them. The foaming was so fast that some of the mixture spilled onto the stove before I could remove the beaker from the heat.

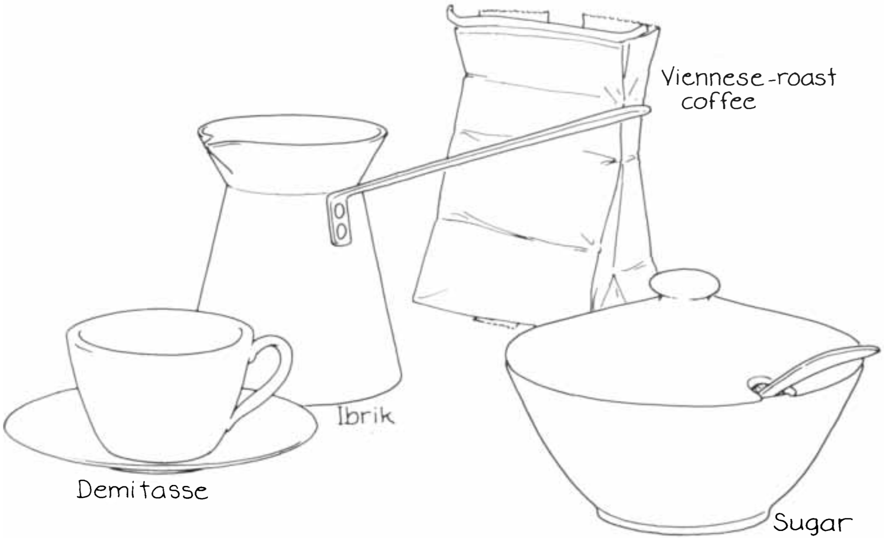
In another preparation I replaced the sugar with an equal amount of honey. I poured the honey into the beaker first, then the water and finally the coffee. When the honey began to heat up, it formed tiny geysers that rose into the layer of water. They were visible only when I shined a flashlight into the beaker, making the difference in the index of refraction between the geysers and the water apparent. If a geyser pinches off from the main layer of honey, it falls back to that layer because honey is denser than water. Often the stream loops or releases drops of honey. Some of the drops remain stationary, apparently because they have been diluted enough to acquire about the same density as that of water.

At this point in my investigation I could answer some of my questions about the brewing of Middle Eastern coffee. The principle of the brewing is to cook the grains fast. If they are overcooked, the drink tastes too bitter or acidic. It is then like regular coffee that has been left simmering in the pot for a long time. The grains for Middle Eastern coffee are finely ground so that they can be rapidly heated and cooked.

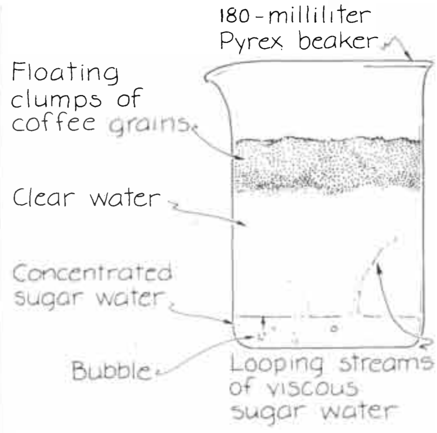
The foam is almost a side effect. The grains are so small that they provide plenty of surface area for adsorbing air. The tiny pockets of air are sites for the nucleation of bubbles. When the coffee and the surrounding water get warm, air dissolved in the water comes out of solution at the nucleation sites, forming bubbles just barely large enough to be seen.

When the hot liquid layer reaches the grains, water vapor expands the bubbles. They pinch off and break free from the nucleating sites and then form the foam of large bubbles that climbs the ibrik. Many of the nucleating sites on the grains remain active because they retain part of the air or vapor after a bubble has left. Moreover, they are now stirred around in the water near the hot bottom of the container. Hence they continue to generate bubbles as long as the surrounding fluid is hot enough to provide more water vapor.

Sugar is added before brewing for several reasons. Heating ensures that the sugar is fully dissolved by the time the coffee is drunk. If the sugar were added after the coffee had been poured into a demitasse, you would have to stir the sugar into the drink to dissolve it. This action would cause the coffee grains to be suspended for much longer. The coffee would be cold before they had set-



The equipment and materials needed to make Middle Eastern coffee



Making the coffee in a beaker

tled out enough to provide a drink that was not gritty.

The sugar also facilitates brewing. The viscous layer of sugar at the bottom of the ibrik holds heat better than water. If the layer consisted of water only, bubbles of water vapor would transfer heat to the coffee layer early in the brewing procedure. With the sugar in place the bubbles and their heat transfer are retarded. Only later does the hot and turbulent layer of sugar water reach the coffee and begin to heat it. The period of heating and cooking of the coffee is sudden and short. It begins when the hot liquid reaches the coffee; it ends in less than a minute when the foam threatens to overflow the container.

Finally, the sugar is beneficial to the brewing process because it helps to stabilize the foam of tiny bubbles. Without the sugar the foam is less likely to survive being poured into the demitasse.

Although the brewing process is fast, it must not be given too little time. When I make enough coffee for two demitasses simultaneously in my ibrik, I often wind up with gritty coffee. No matter how long I wait large grains fail to settle. The reason is that they still carry enough adsorbed air to remain buoyant. They may even remain clumped together. Since the ibrik has too much fluid, I must take it off the stove prematurely to keep foam from overflowing. As a result the time from when the hot, turbulent sugar layer reaches the coffee to when the coffee is poured is too short for all the adsorbed air to be eliminated. From experience I have learned to prepare only one demitasse of coffee at a time. If it is necessary to make a larger amount in one preparation, I must hold the ibrik off the heat source to control the foam and to cook the coffee long enough.

My next set of questions has to do with what is going on in the demitasse as the grains of coffee sink to the bottom. Each grain sinks at a speed that is approximately constant. Since the sedimentation rate of suspended particles is an important datum in both science and technology, much work has been done to develop a theory of how fast particles settle in water.

If the particles are smaller than about .06 millimeter in diameter, they fall according to Stokes's law, named after the British physicist and mathematician G. G. Stokes. The speed of a small particle depends on two factors: the effective weight of the particle and the drag applied to it by the fluid. The effective weight is the difference between the particle's true weight and the buoyancy imparted to it by the surrounding fluid. This difference is usually described in terms of the density of the particle and the fluid. Since a coffee grain is denser than water (and denser than a mixture of sugar and water), the buoyancy imparted to it is too weak to keep it from falling. Its effective weight is a downward force smaller than its true weight.

Countering this force is an upward force generated by drag. According to Stokes's law, the drag depends on the diameter of the particle, the speed of the particle and the viscosity of the fluid. When the particle begins to descend, its speed is small and so is the drag. The particle steadily accelerates and soon is moving fast enough for the upward drag force to match the downward force of the particle's effective weight. Thereafter the particle no longer accelerates, but it continues to fall at a constant speed. This terminal speed depends on the size of the particle. Smaller particles have lower terminal speeds and therefore take longer to settle.

The value of the terminal speed also depends on the viscosity of the fluid. Since the sugar increases the viscosity of the coffee, all the particles have lower terminal speeds than they would have without the sugar.

Unfortunately for drinkers of Middle Eastern coffee, Stokes's law applies only to spherical particles. The coffee grains small enough to be affected by the law are irregular in shape and therefore encounter a drag force that is difficult to calculate. Nevertheless, the overall dependence on particle size generally holds: smaller grains fall slower.

Many of the grains in Middle Eastern coffee are too large to fall according to Stokes's law, which predicts a drag force that depends on the first power of the particle's speed through the fluid. The larger particles get a drag force that depends on the square of the speed. In addition the equation for the force includes an experimentally determined drag coefficient, which takes into account the effect of the particle's irregular shape.

Another factor is that if there are many particles falling through a certain volume of fluid, they change the situation in two ways. They increase the effective density of the fluid, thereby decreasing the effective weight of a particle and its settling rate. The particles also interact as they pass one another on the way down. In spite of this added complication the same rule usually holds: smaller particles fall slower.

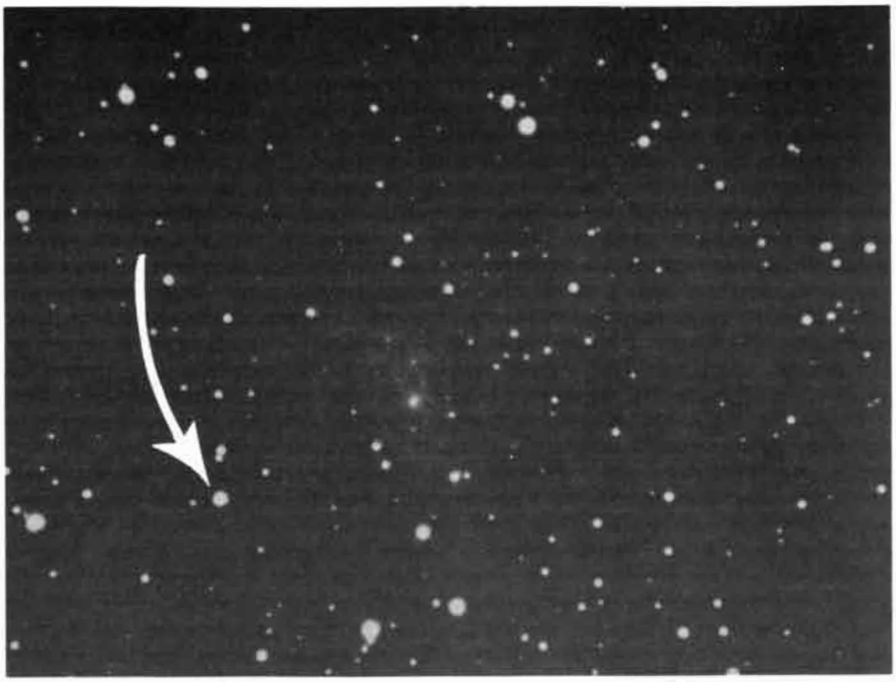
To demonstrate the rule I poured a small amount of freshly brewed Middle Eastern coffee into a tall, transparent jar containing water. I had left the water undisturbed for an hour so that it was still. I added freshly brewed coffee because I wanted grains that had lost most of their adsorbed air. Unbrewed grains would have tended to float.

With the aid of a flashlight I monitored the descent of the particles. Large ones sank quickly to the bottom at a speed of about one centimeter per second. Smaller grains were much slower. Some of the smallest ones appeared to take 10 minutes or more to reach the bottom.

The settling rate of particles in a demitasse is somewhat different from what I saw in my demonstration. The number of particles per unit of volume is large enough to make it likely that the particles interact as they fall. More important, the fluid is hot enough to create circulation cells. Hot fluid from near the bottom rises to the top, cools by evaporation and then descends. The circulation forms what are called Bénard cells, after the French scientist Henri Bénard. In some fluids and under certain heating conditions the cells create a stable geometric pattern on the top surface.

In my coffee the pattern is chaotic and constantly changing. It is best seen when the coffee lacks foam, since foam not only covers the surface but also alters the evaporation rate. I rig a light that reflects off the surface, revealing that over the region of ascending hot fluid tiny drops are suspended just above the surface. They condense from the water vapor when it cools in the air just above the surface. No drops are visible over the regions of descending fluid. Since the drops reflect light, the regions of ascending fluid look slightly white while the regions of descending fluid are the dark color of the coffee. Bénard cells in a cup of coffee with no grains are merely a novelty. In Middle Eastern coffee the circulation in Bénard cells delays the settling of the grains.

The result of all this physics is that one would do well to strike a balance when drinking Middle Eastern coffee, waiting long enough for the larger particles of coffee to settle to the bottom of the demitasse and for the Bénard circulation to decrease but not so long that the drink gets unpleasantly cool. The foam on the coffee makes the compromise easier by insulating the coffee with the vapor trapped in the bubbles, thereby decreasing the cooling rate.



An example of the interesting things that might turn up unexpectedly while you are casually sweeping the deep sky is this supernova in NGC 6946, in a portion of a photograph taken by Hubert Entrop on November 4, 1980, with his Questar. This was 7 days after the official discovery of the nova, as described in Sky & Telescope *in the January, 1981, issue.*

BIBLIOGRAPHY

Readers interested in further explanation of the subjects covered by the articles in this issue may find the following lists of publications helpful.

METAMAGICAL THEMAS

RECURSIVE FUNCTIONS. Rózsa Péter. Academic Press, 1967.

ANATOMY OF LISP. John Allen. McGraw-Hill Book Company, 1978.

ARTIFICIAL INTELLIGENCE PROGRAMMING. Eugene Charniak, C. K. Riesbeck and Drew V. McDermott. Lawrence Erlbaum Associates, Inc., 1980.

HISTORY OF LISP. John McCarthy in *History of Programming Languages,* edited by Richard L. Wexelblat. Academic Press, 1980.

A NUCLEAR-WEAPON-FREE ZONE IN EUROPE

NUCLEAR-WEAPON-FREE ZONES. Report from Vantage Conference 5, Muscatine, Iowa, October 7–9, 1975. The Stanley Foundation, 1975.

COMPREHENSIVE STUDY OF THE QUESTION OF NUCLEAR-WEAPON-FREE ZONES IN ALL ITS ASPECTS. Special Report of the UN Conference of the Committee on Disarmament. United Nations, 1976.

AMERICAN FOREIGN POLICY AND NUCLEAR-WEAPON-FREE ZONES: ISSUES AND OPTIONS. Joseph M. Grieco in *Potomac Review,* Vol. 7, pages 12–23; Winter, 1976.

ARMS CONTROL: THE REGIONAL APPROACH. Ronald G. Purver. Center for International Relations, Queens University, Kingston, 1981.

STUDY ON ALL THE ASPECTS OF REGIONAL DISARMAMENT. United Nations Centre for Disarmament. Report to the Secretary General, United Nations, 1981.

COMMON SECURITY: A BLUEPRINT FOR SURVIVAL. Independent Commission on Disarmament and Security Issues. Simon and Schuster, 1982.

SILICON MICROMECHANICAL DEVICES

SPECIAL ISSUE ON THREE-DIMENSIONAL DEVICE STRUCTURES. *IEEE Transactions on Electron Devices,* Vol. ED-25, No. 10; October, 1978.

SPECIAL ISSUE ON SOLID-STATE SENSORS, ACTUATORS, AND INTERFACE ELECTRONICS. *IEEE Transactions on Electron Devices,* Vol. ED-26, No. 12, December, 1979.

SILICON AS A MECHANICAL MATERIAL. Kurt E. Petersen in *Proceedings of the IEEE,* Vol. 70, No. 5, pages 420–457; May, 1980.

SPECIAL ISSUE ON SOLID-STATE SENSORS, ACTUATORS, AND INTERFACE ELECTRONICS. *IEEE Transactions on Electron Devices,* Vol. ED-29, No. 1; January, 1982.

THE STRUCTURE OF QUARKS AND LEPTONS

THE KEY TO THE UNIVERSE: A REPORT ON THE NEW PHYSICS. Nigel Calder. British Broadcasting Corporation, 1967.

GAUGE THEORIES OF THE FORCES BETWEEN ELEMENTARY PARTICLES. Gerard 't Hooft in *Scientific American,* Vol. 242, No. 6, pages 104–138; June, 1980.

A UNIFIED THEORY OF ELEMENTARY PARTICLES AND FORCES. Howard Georgi in *Scientific American,* Vol. 242, No. 4, pages 48–63; April, 1981.

HOT SPRINGS ON THE OCEAN FLOOR

THE EVOLUTION OF SEDIMENTARY ROCKS. Robert M. Garrels and Fred T. Mackenzie. W. W. Norton and Co., Inc., 1971.

SEAWATER-BASALT INTERACTION AT 200° C AND 500 BARS: IMPLICATIONS FOR ORIGIN OF SEA-FLOOR HEAVY-METAL DEPOSITS AND REGULATION OF SEAWATER CHEMISTRY. James L. Bischoff and Frank W. Dickson in *Earth and Planetary Science Letters,* Vol. 25, No. 3, pages 385–397; April, 1975.

THE MECHANISMS OF HEAT TRANSFER THROUGH THE FLOOR OF THE INDIAN OCEAN. Roger N. Anderson, Marcus G. Langseth and John G. Sclater in *Journal of Geophysical Research,* Vol. 82, No. 23, pages 3391–3409; August 10, 1977.

CHEMISTRY OF HOT SPRINGS ON THE EAST PACIFIC RISE AND THEIR EFFLUENT DISPERSAL. J. M. Edmond, K. L. Von Damm, R. E. McDuff and C. I. Measures in *Nature,* Vol. 297, No. 5863, pages 187–191; May 20, 1982.

A WINDOW ON THE SLEEPING BRAIN

THE FUNCTIONS OF SLEEP. Edited by René Drucker-Colin, Mario Shkurovich and M. B. Sterman. Academic Press, Inc., 1979.

SLEEP DISORDERS: INTERSECTIONS OF BASIC AND CLINICAL RESEARCH. Edited by M. H. Chase. Spectrum Publications, Inc., 1982.

DIFFERENT BEHAVIORS DURING PARADOXICAL SLEEP WITHOUT ATONIA DEPEND ON PONTINE LESION SITE. J. C. Hendricks, A. R. Morrison and G. L. Mann in *Brain Research,* Vol. 239, pages 81–105; 1982.

EARLY FARMERS OF THE NORTH EUROPEAN PLAIN

EUROPEAN PREHISTORY. Sarunas Milisauskas. Academic Press, 1978.

THE PREHISTORY OF EUROPE. Patricia Phillips. Indiana University Press, 1980.

EARLY NEOLITHIC SITES AT BRZEŚĆ KUJAWSKI, POLAND: PRELIMINARY REPORT ON THE 1976–1979 EXCAVATIONS. Ryszard Grygiel and Peter Bogucki in *Journal of Field Archaeology,* Vol. 8, No. 1, pages 9–27; 1981.

EARLY EUROPEAN AGRICULTURE. M. R. Jarman, G. N. Bailey and H. N. Jarman. Cambridge University Press, 1982.

EARLY NEOLITHIC SUBSISTENCE AND SETTLEMENT IN THE POLISH LOWLANDS. Peter I. Bogucki. British Archaeological Reports, Oxford, 1982.

CHEMICAL SIGNALS OF SOCIAL AMOEBAE

THE CELLULAR SLIME MOLDS. John Tyler Bonner. Princeton University Press, 1967.

THE DEVELOPMENT OF *DICTYOSTELIUM DISCOIDEUM.* Edited by W. F. Loomis. Academic Press, 1982.

CHEMICAL IDENTITY OF THE ACRASIN OF THE CELLULAR SLIME MOLD, *POLYSPHONDYLIUM VIOLACEUM.* O. Shimomura, H. L. B. Suthers and J. T. Bonner in *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 79, No. 23, pages 7367–7379; December, 1982.

INTUITIVE PHYSICS

MEDIEVAL AND EARLY MODERN SCIENCE: VOL. 2. Alistair Crombie. Doubleday & Company, Inc., 1959.

CURVILINEAR MOTION IN THE ABSENCE OF EXTERNAL FORCES: NAÏVE BELIEFS ABOUT THE MOTION OF OBJECTS. Michael McCloskey, Alfonso Caramazza and Bert Green in *Science,* Vol. 210, No. 4474, pages 1139–1141; December 5, 1980.

STUDENTS' PRECONCEPTIONS IN INTRODUCTORY MECHANICS. John Clement in *American Journal of Physics,* Vol. 50, No. 1, pages 66–71; January, 1982.

THE AMATEUR SCIENTIST

OBSERVATIONS OF AN EARLY MORNING CUP OF COFFEE. Vincent J. Schaefer in *American Scientist,* Vol. 59, No. 5; September–October, 1971.

GRAIN SHAPE EFFECTS ON SETTLING RATES. P. D. Komar and C. E. Reimers in *Journal of Geology,* Vol. 86, No. 2, pages 193–209; March, 1978.

SCIENCE AND THE CITIZEN

A Social Price

Amounts spent on arms and armies continued to increase last year, reaching a total of $600 billion. The total is given by Ruth Leger Sivard in "World Military and Social Expenditures 1982," the latest in a series of reports begun 15 years ago by the U.S. Arms Control and Disarmament Agency and now privately sponsored. The report aims "to provide an annual accounting of the use of world resources for social and for military purposes, and an objective basis for assessing relative priorities." The contention of the report is that "relative to the economy, the burden [of military expenditures] is growing larger." In other words, the world is paying a social price for the arms race.

From 1960 through 1980 world military expenditures increased by 441 percent. If the amounts are converted into 1979 dollars to remove the effects of inflation, the increase is 66 percent. Over the same period the sum of the gross national products of all countries (adjusted for inflation) increased by 146 percent. The average annual increase in military spending was 3.3 percent, and the average annual increase in G.N.P. was 7.3 percent. During most of the past 20 years, therefore, armaments claimed a decreasing share of the world's wealth. Between 1979 and 1980, however, military spending corrected for inflation went up by 3.6 percent and the average G.N.P. by only 2.6 percent.

The perspective is slightly different if increases in population, which offset part of the gain in G.N.P., are taken into account. The per capita G.N.P. calculated in 1979 dollars increased by 68 percent in the years from 1960 through 1980, or an average of 3.4 percent per year, much less than the average 7.3 percent increase in total G.N.P. Between 1979 and 1980 the per capita G.N.P., corrected for inflation, rose by less than 1 percent, compared with the 3.6 percent increase in military spending.

When the nations of the world are broken down into the two categories of developed and developing countries, it becomes apparent that the developing countries have paid a comparatively heavy social price for military spending. The military expenditures of the developed countries (in 1979 dollars) have increased by 44 percent in the past 20 years, whereas those of the developing countries have increased by 279 percent. The population of the developed countries increased by 20 percent and that of the developing countries by 58 percent. Because of the difference in population growth, the increases in per capita G.N.P. were 102 percent and 91 percent respectively, even though the G.N.P. of the developing countries increased by 219 percent and that of the developed countries by 130 percent.

The U.S. and the U.S.S.R. spend half of the world's military budget. "Both countries," the report comments, "the U.S.S.R. especially, have used a larger-than-average share of an expanding economic base to support the military effort." Whereas 4.5 percent of the world G.N.P. went for military expenditures in 1979, the U.S. spent 5.2 of its G.N.P. on the military and the U.S.S.R. spent 10.7 percent (although there are uncertainties in estimates of Russian military spending). The U.S. and the U.S.S.R. are also currently investing at least twice as much in military research and development as they are in nonmilitary research and development.

The report concludes with a breakdown of the military and social expenditures of 141 nations in 1979. Each nation is assigned a social-economic rank based on per capita G.N.P. and various indicators of the quality of education and health care, such as the population per physician. The U.S. is ranked seventh, after the Scandinavian countries, France and Australia. The U.S.S.R. is 23rd on the list.

W

With the discovery of the *W* particle a major goal of experimental elementary-particle physics has been achieved. The *W* belongs to the class of particles known as intermediate vector bosons, or weakons. According to the now-standard theory linking electromagnetism and the weak nuclear force, there are three such particles: two of them are electrically charged (the W^+ and W^-) and one is neutral (Z^0). In the unified electroweak theory the bosons act as intermediaries between weakly interacting particles, just as the photon serves as the carrier of the electromagnetic force between electrically charged particles. In contrast to the massless photon, however, the *W* and *Z* bosons have a prodigious mass: more than 80 times that of the proton or the neutron. Indeed, the large mass of the weakons long put them out of reach of particle accelerators [see "The Search for Intermediate Vector Bosons," by David B. Cline, Carlo Rubbia and Simon van der Meer; SCIENTIFIC AMERICAN, March, 1982].

The first instrument with enough energy to create such particles was built almost two years ago by modifying an accelerator at CERN, the European Laboratory for Particle Physics in Geneva. The CERN machine, which was designed primarily to look for intermediate vector bosons, brings counterrotating beams of protons and antiprotons into collision at a total energy of 540 GeV (billions of electron volts). After a series of preliminary runs in 1981, followed by an unexpected delay caused by an accident in one of the accelerator's two main particle detectors, the search resumed in earnest late last year. During a 30-day period in November and December approximately a billion high-energy proton-antiproton collisions were recorded for subsequent computer-aided analysis. By January physicists at CERN were able to announce that they had at last detected evidence of the fleeting existence of *W* particles in the aftermath of a small fraction of the collision events.

Two experimental groups were responsible for the investigation. One group, headed by Carlo Rubbia of CERN and Harvard University, included 126 physicists from 13 universities and research centers in Europe and the U.S. The other, headed by Pierre Darriulat of CERN, numbered 51 physicists from six European laboratories.

The first group, working with a large, multipurpose particle detector named UA-1 (for Underground Area One), has so far identified at least five events "revealing the expected signature of the *W* boson." According to an account of the group's findings submitted to *Physics Letters,* two independent analytic techniques were employed to distinguish the events of interest from the large number of other events. In one approach the group looked for the track of a high-energy electron or positron (the electron's antiparticle) emitted from the collision zone at a large angle with respect to the axis of the colliding beams. In the other approach the group looked for evidence of a significant amount of "missing energy" attributable to the emission of a neutrino from the collision zone. (Because neutrinos have no electric charge, they cannot be tracked by the electromagnetic detection system.)

In each case the experimenters were left with five unambiguous *W* candidates and one probable candidate. Comparison of the two sets of data showed that both methods of analysis had identified the same events. As the experimenters note in their report, "the simultaneous presence of an electron and (one) neutrino of approximately equal and opposite momenta in the transverse direction... suggests the presence of a two-body decay." The decay is interpreted as the spontaneous disintegration of a short-lived W^- particle into an electron and an antineutrino or of an equally short-lived W^+ particle into a pos-


DATSUN MAXIMA
OBVIOUS LUXURY...SURPRISING PERFORMANCE
Introducing the luxury sedan that was born to perform. The magnificent Datsun Maxima boasts a fuel-injected 2.4 liter six cylinder engine that merges surprising power with unsurpassed elegance.
Inside, virtually every conceivable creature comfort: cruise control on both the 5-speed and automatic, eight speaker stereo including cassette deck with Dolby, air conditioning, power windows, mirrors and door locks. A voice will tell you when your "Right door is open"—one of six vocal reminders. All this is standard.
The optional sunroof is power driven. The optional leather package includes a new digital control panel that monitors the engine's functions electronically.
The Datsun Maxima is available in four-door sedan or wagon, gas or diesel. Compare it feature by feature with any other car in its luxury class. Compare the performance. Then compare the price. Most importantly of all, remember it's built by the Nissan Motor Company, Ltd.—whose name stands for quality world-wide.
DATSUN
WE ARE DRIVEN
NISSAN
Product of NISSAN
© 1983 NISSAN MOTOR CORPORATION IN USA

The world's densest computer is now the heart of the world's most 32-bit computer.

Our 32-bit CPU.

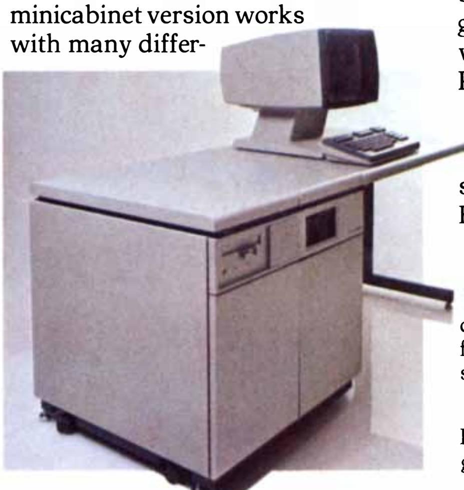
From time to time, miracles of technology come along to make previously impossible tasks not only possible, but easy. That tiny 450,000-transistor integrated circuit is one of those technological miracles.

Hewlett-Packard didn't develop it just to break the record for most transistors on a chip, but to put on an engineer's or scientist's desk a computer so powerful that it can do the work of mainframes costing four times as much.

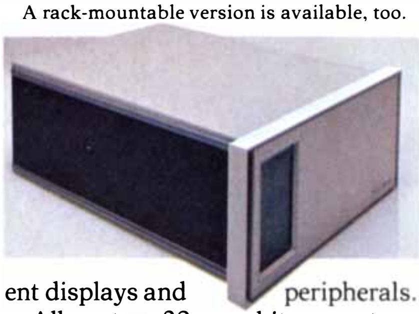
32-bit computers for 32-bit applications.

The new HP 9000 computer based on this and four other 'superchips' can handle formidable engineering and scientific problems. The scientist solving complex systems of equations, the mechanical engineer doing finite element analysis or three-dimensional modeling, the electrical engineer analyzing complex circuits or designing very large-scale integrated circuits—these are the kinds of technical people and problems the HP 9000 family is designed for.

It comes in three versions. The integrated workstation is complete with keyboard, color or monochromatic graphics display, fixed and flexible disc drives, and printer. For systems manufacturers, there's a rack-mountable box. And for a variety of single-user and multi-user applications, the minicabinet version works with many different displays and peripherals.



As a minicabinet, it can handle multiple users.



A rack-mountable version is available, too.

All are true 32-bit computers, with 32-bit CPUs, memories, and data paths. And the multi-CPU architecture lets you nearly double or triple your processing power at any time by adding one or two CPU boards. Without increasing the computer's size.

Two operating systems are better than one.

The integrated workstation is available with a choice of operating systems. One is HP's highly evolved, high-performance Enhanced BASIC, augmented with 3-D graphics and a software innovation called a run-time compiler. This substantially increases program execution speed, while retaining an interactive development environment.

The other operating system, called HP-UX, is a fully supported, extended version of the popular UNIX.® HP-UX, available on all HP 9000s, adds virtual memory, graphics, data base management, data communications, and enhanced file capability to the basic UNIX 'shell.' High-level programming languages available with HP-UX are FORTRAN 77, Pascal and C.

Software, and plenty of it.

Much of the vast range of existing software written in HP BASIC, FORTRAN 77, Pascal and C is transportable to the HP 9000. HP will also be offering proprietary software packages emphasizing computer-aided design and engineering. These will tie the HP 9000 into HP's Manufacturer's Productivity Network (MPN). Third-party software suppliers will be providing many of the most widely used CAE packages for 32-bit computer systems. And both HP 9000 operating

The 32-bit CPU chip is bonded to the finstrate which doubles as a signal carrier and heat sink.

Up to three CPU boards and three Input/Output Processors can fit into a single HP 9000.

TCG-207

chip
affordable

systems offer extensive program development tools.

You also get a choice of communication tools. The HP 9000 is currently compatible with Ethernet,™ and with HP's Shared Resource Manager (SRM) which lets clusters of HP 9000 and 16-bit desktop computers share data and use common peripherals. Links to central computers are also available. And in late 1983, HP will offer local area networks based on the IEEE-802 standard.

New technology from the silicon up.

The five superchips that make the HP 9000 possible are the 32-bit CPU, which can execute a million instructions per second; an eight-channel Input/Output processor (IOP); a random-access memory chip capable of storing 128K bits of data; a memory controller that 'heals' up to 32 bad memory locations; and an 18-megahertz clock.

Hewlett-Packard's advanced NMOS-III process makes it possible to put 450,000 transistors on a chip only 0.4 centimeters square. This tremendous density of electronic components could have required an expensive and elaborate cooling system.

Instead, HP engineers developed a new mounting structure called a finstrate, a copper-cored circuit board, which acts as both cooling fin and substrate. The finstrates containing the CPU, IOP, memory, and clock chips are housed in a lunchpail-sized module.

One user, one mainframe.

Clearly the trend in engineering and scientific computation is away from large machines shared by multiple users and towards networks of powerful personal workstations, sharing peripherals and data bases. The reason is compelling. An engineer or scientist in personal control of an HP 9000 can solve so many more problems more easily that the increased productivity alone makes the cost of individual computers easy to justify.

For complete information about this powerful breakthrough in 32-bit computing, contact the local HP sales office listed in your telephone directory. Ask a Technical Computer Specialist for a demonstration. Or write to Pete Hamilton, Dept. 41151, Hewlett-Packard, 3404 East Harmony Road, Fort Collins, CO 80525.

Full-color or monochromatic display. 3-D graphics are available.

Eight soft keys play an important role in the menu-driven operation.

Built-in thermal printer produces graphics and alphanumeric hard copy.

Flexible disc drive.

Optional 10-Mbyte Winchester disc.

HEWLETT PACKARD

UNIX is a trademark of Bell Laboratories. Ethernet is a trademark of Xerox Corporation.

itron and a neutrino. The mass of the W particles implied by measurements of the decay products is 81 ± 5 GeV, which the group points out is "in excellent agreement" with the prediction of the electroweak theory.

The second group, working with a more specialized and somewhat simpler detector named UA-2, announced in January that it had found four events "that are consistent with a W signature." The group added, however, that "more work is needed to confirm this preliminary UA-2 result." Both groups will get a chance to substantiate their early findings when the CERN proton-antiproton collider starts a new series of observations this month. High on the agenda will be the search for the Z^0, which ought to materialize about 10 times less frequently than the W particles.

Unpolishing the Apple

In the academic year 1980–81 there were about 140,000 students at the nine campuses of the University of California. Of this number 22 students were enrolled in programs leading to certification as a teacher of mathematics in elementary or secondary school. There were 47 students in courses leading to certification as a science teacher. In the same academic year there were about 300,000 students attending the 19 campuses of the California State University system; 75 were in training to be mathematics teachers and 127 to be science teachers. The numbers are cited by James Guthrie and Ami Zusman of the University of California School of Education in a report on the training of mathematics and science teachers in California; the report was published recently by the university's Institute of Governmental Studies.

One reason few students seem to be interested in becoming mathematics or science teachers is the inadequacy of teachers' salaries. The average salary for a beginning teacher in the San Francisco Bay area is $12,680. University graduates with the qualifications needed to become a certified teacher can get jobs in industry at a starting salary of about $20,000. In addition to reducing the number of future teachers, the higher pay offered by industry is depleting the current ranks of elementary and high school teachers. Guthrie and Zusman cite a survey done in the high school districts around Palo Alto. The area is part of the region known as Silicon Valley and has a high concentration of businesses involved with advanced technology. The survey found that six of the eight school districts in the area were losing mathematics and science teachers to better-paid jobs in industry.

Ironically, it is in areas where industry is based largely on new technology that the need for excellence in science education is most acute. Jobs in such industries require a comparatively high level of scientific and technical knowledge. Guthrie and Zusman estimate that between 1980 and 1990 about 40 percent of the new jobs in California will be in businesses that employ new technology. Thus California is faced with an increasing demand for technically trained workers and a decreasing supply of teachers to educate them.

The disjunction between supply and demand in science education may be greater in California than it is elsewhere in the U.S., but the problem is of national scope. In 1982 the National Science Board of the National Science Foundation established a Commission on Precollege Education in Mathematics, Science and Technology. In its first formal report the commission concludes that the national system of science education is failing to give workers adequate technical skills. Indeed, the commission concludes that this is one of two major functions of science education that are not being fulfilled. The other is providing the general public with sufficient knowledge to enable citizens to make informed decisions on matters related to science and technology. A third function, that of training professional scientists, mathematicians and engineers, is being carried out better than the other two, the commission states.

The shortage of qualified teachers was found to be widespread. In 1981, 43 states out of the 45 that responded to a survey said they had a shortage of mathematics teachers; 42 of the 45 had a shortage of physics teachers. From 1971 through 1980 the number of students preparing to become teachers decreased by a factor of three in science and by a factor of four in mathematics. Moreover, of those who had trained to teach between 1971 and 1980 only half actually entered the teaching profession. A fourth of those who were teaching said they planned to leave for other jobs in the near future.

In an attempt to reverse the trend the Reagan Administration has proposed appropriating $200 million for the training of mathematics and science teachers. The funds would enable states to make grants of up to $5,000 to teachers who want to take up mathematics or science teaching. Those eligible for the grants might include unemployed teachers, teachers in other fields of study and teachers of mathematics or science who have no certification in those subjects.

According to the Commission on Precollege Education, only a minority of high school students in the U.S. now receive an adequate scientific education. Only a third of the 21,000 American high schools offer a course in calculus. Fewer than a third offer a course in physics taught by a qualified teacher. The report concludes: "The educational system may actually have carried out [the education of skilled workers and the general public] better 20 years ago: the proportion of public high school students enrolled in science courses has declined since that time."

Oncogenetic Switch

Human cells harbor genes that have a function in normal metabolism but that can be triggered to promote cancerous growth. The discovery of such oncogenes has provided both a conceptual framework and a set of powerful tools for investigating the molecular genetics of cancer. A number of lines of research have now begun to make sense of what had seemed to be disparate observations concerning the causes and manifestations of malignancy.

The essential characteristic of a cancer is the uncontrolled proliferation of cells. The product of one family of oncogenes is a protein kinase, an enzyme that adds phosphate groups to proteins; the oncogene kinases are unusual in that they phosphorylate the amino acid tyrosine, whereas most kinases phosphorylate serine or threonine. The receptors for insulin and some other cellular growth factors can also phosphorylate tyrosine. The receptors do so, however, only in the presence of the appropriate growth factor, whereas the oncogene enzymes throw a switch that commits cells to uncontrolled proliferation. The phosphorylation of tyrosine may therefore be central to growth control, and aberrant phosphorylation may be one step in carcinogenesis.

Identifying the proteins encoded by oncogenes and determining what the proteins do in the cell is one objective of current research. Another compelling objective is to learn what activates an oncogene, changing a normal gene into a cancer gene. Perhaps the carcinogenic version of the gene has undergone mutation or reshuffling, so that it specifies a different protein: an enzyme or regulatory molecule that is either more or less efficient than the normal one, say, or an enzyme with an altered substrate specificity. Perhaps, on the other hand, the protein-coding region of the gene is not changed but the level of its expression is enhanced, so that too much of a normally benign protein is synthesized. In the past few months some striking results from two laboratories have pinpointed a possible mechanism of oncogene activation that would be compatible with either the mutational or the dosage hypothesis or with both.

A number of cancers, particularly those affecting various blood and lymphatic cells, are associated with chromosomal abnormalities. One kind of abnormality is a reciprocal translocation, in which segments of two chromosomes are interchanged. Several investigators, including John Cairns of the Harvard School of Public Health, George Klein of the Karolinska Institute and Janet D. Rowley of the University of Chicago School of Medicine, have pointed out that a translocation might move a potential oncogene into a region where DNA is being actively transcribed and thereby enhance its expression. Last year several oncogenes were mapped to specific chromosomal segments that are indeed subject to translocation.

The gene designated c-*myc* is the cellular analogue of a viral oncogene originally recognized for its ability to induce tumors in chickens and other birds. The gene is near one end of human chromosome 8. In a chromosomal abnormality associated with Burkitt's lymphoma, a cancer of antibody-producing *B* lymphocytes, the end of chromosome 8 is translocated. Most frequently it moves to chromosome 14, to a region carrying genes for the heavy chain of antibody molecules.

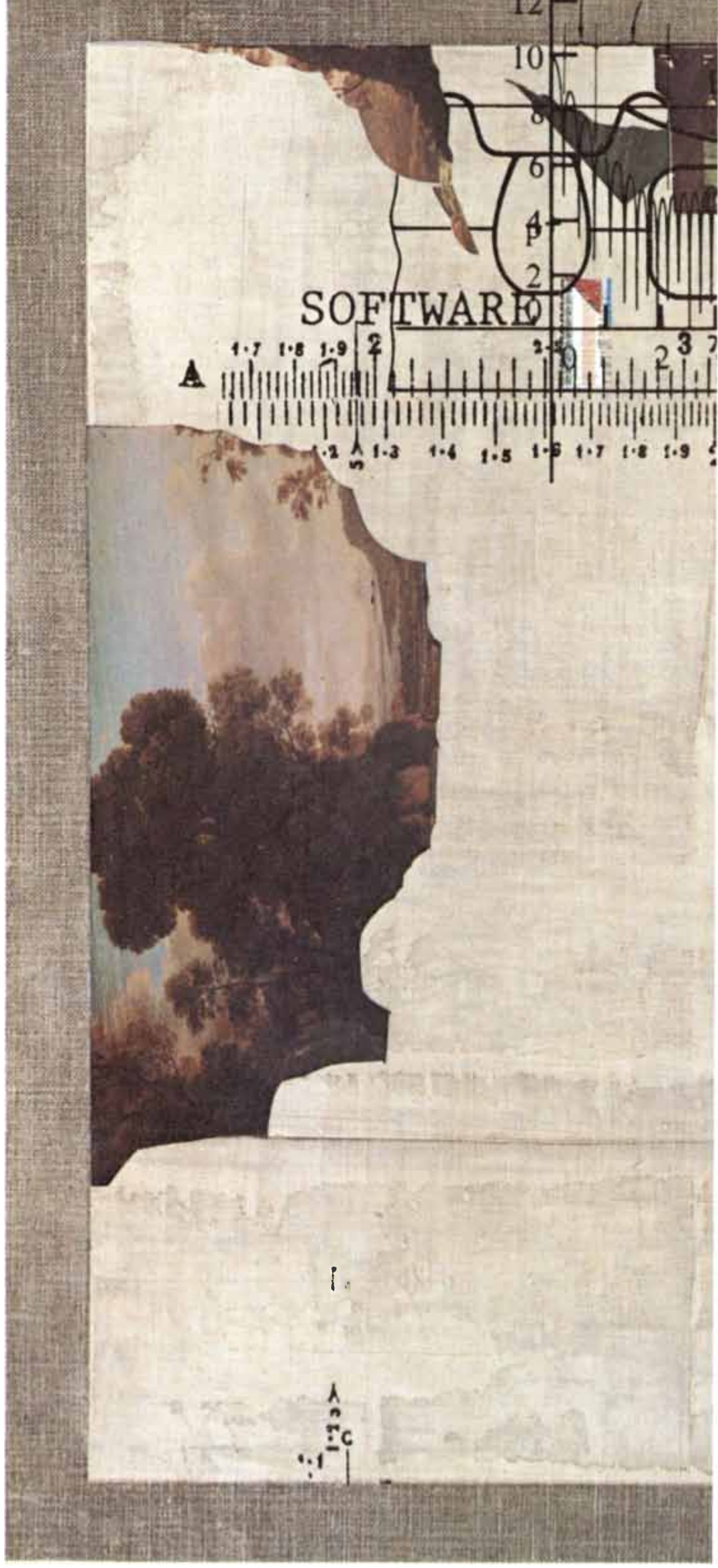
In lymphocytes the antibody-gene loci are particularly active ones; in the course of the immune response they are shuffled in various ways and are transcribed into RNA that is translated to make specific antibodies of various classes. Now two groups, headed by Philip Leder of the Harvard Medical School and by Carlo M. Croce of the Wistar Institute of Anatomy and Biology, have reported in *Proceedings of the National Academy of Sciences* that the locus of c-*myc* on chromosome 8 is precisely at the "breakpoint" of the 8-to-14 translocation. In several Burkitt's-lymphoma cell lines c-*myc* apparently ends up associated with a particular heavy-chain antibody gene on chromosome 14.

The implication is that translocation to the heavy-chain locus brings c-*myc* under the influence of a "promoter" region that ordinarily mediates the transcription of a particular heavy-chain gene; in Burkitt's lymphoma the region may instead promote the active transcription of c-*myc* and thereby enhance its expression. (So far Leder's group has not observed a dramatic increase in c-*myc* RNA in lymphoma cell lines; Croce's group reports a significant increase.) An alternative possibility is that in the course of the translocation c-*myc* is somehow rearranged so that it encodes a different product. Further analysis of the translocated gene, together with more study of c-*myc*'s protein product (of which little is yet known), should tell whether mutation or dosage is at work in this case.

Spectral Memory

The presence or absence of a mark at a predetermined position on a sheet of paper encodes one bit of information. If the mark is allowed to vary in color as well as position, the density of the infor-

What do IBM people think about?



Creating better computers takes a lot of different thinking from a lot of different people.

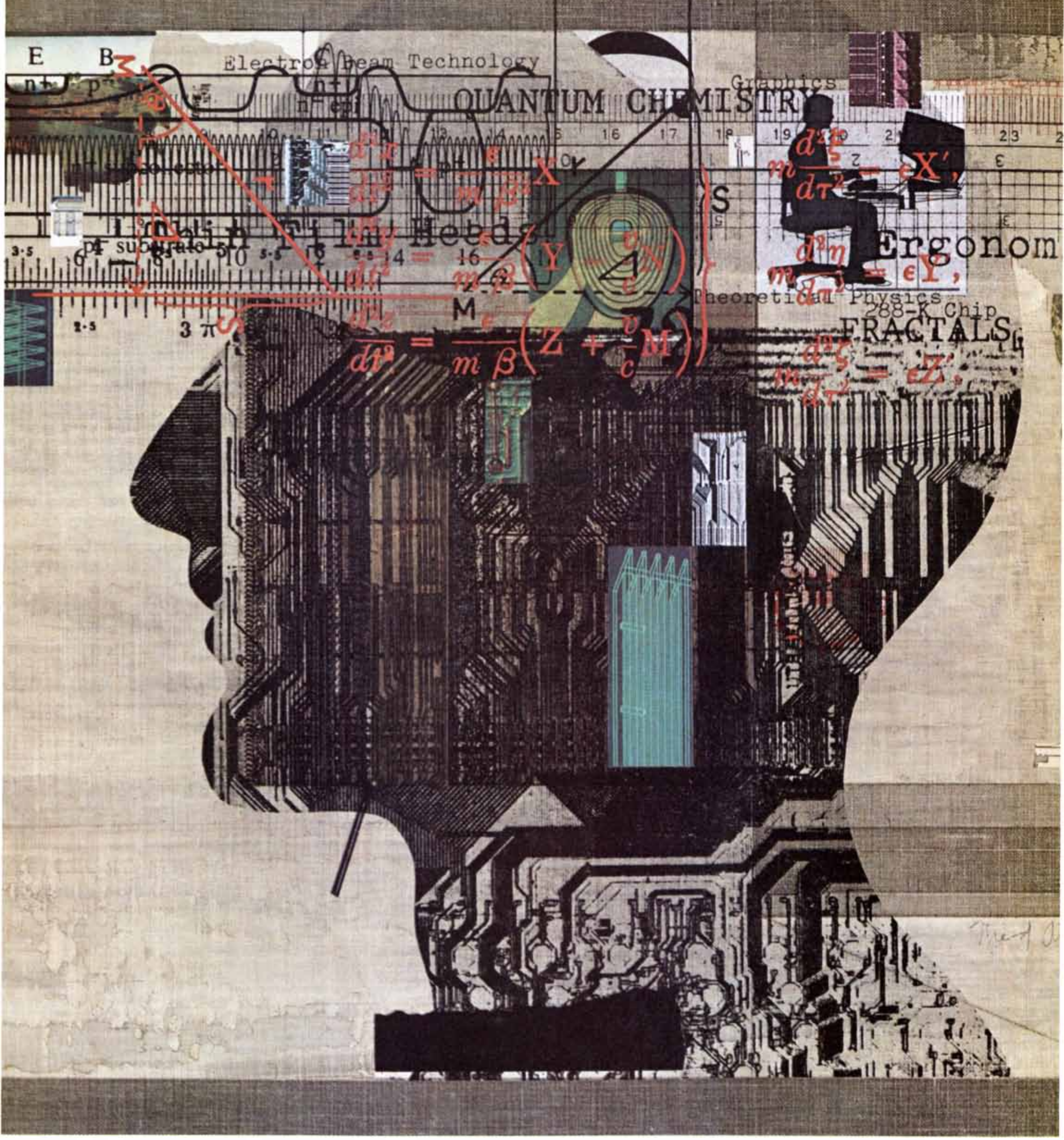
At IBM, our scientists and engineers represent a wide range of disciplines.

For example, IBM chemists are exploring the structure of ceramics and developing special polymers to make computers more compact.

IBM physicists are studying how short light pulses travel through optical fibers to build faster computer systems.

There are IBM psychologists and human factors specialists who study how people interact with machines, so we can make them friendlier and easier to use.

In laboratories and plants all over the world, thousands of



IBM people are putting their minds to work on questions of science and technology. They're all part of a 50-year commitment to research and development—a continuing commitment that has totalled more than $8 billion in the last six years alone.

Because finding new ways to improve our computers is one thing that's always on our minds. IBM®

Some of the most powerful forces in communications are at work at CSD

Systems Engineers
Hardware Engineers
Software Engineers
ATE Engineers
R/M Engineers

We want engineers who can see the total picture, whether they're working on independent R&D or leading an entire department. People who can translate customer needs into creative solutions, identify technologies, make the right design decisions and lead in high visibility positions are going to succeed here.

Engineers like you have found unlimited growth at CSD. Some of the leading minds in communications engineering are working here today on projects like those described in this ad. Our involvement goes across the entire communications spectrum from DC to daylight.

We are now growing so rapidly that **if you meet the criteria described in this ad, we can create the right opportunity for you.**

Among our broad range of projects:

TRANSMISSION SYSTEMS

We're engineering integrated radio systems using digital wide band processing. New technologies include unique fiber optic applications and systems operating in the VHF to EHF range.

MILSATCOM

Our emergent-technology EHF band systems involve AJ protection, rapid configuration and interoperability with satellite networks. Key systems elements include EHF antennas, feeds and high performance synthesizers.

DIGITAL PROCESSING SYSTEMS

These high performance systems cover the range of audio, ELF, VLF, LF and HF. We're developing algorithms with critical timing and throughput requirements, as well as applying high speed data processing in advanced digital communications products that include digital modems, voice and receivers, and systems controllers.

NETWORKING

We're pioneering totally integrated multinetwork interface systems pushing the state-of-the-art. Our extensive involvement in networking includes new concepts in such areas as message, packet and digital switching.

SECURE SYSTEMS

The security/intelligence community looks to us for leadership in systems knowledge. We're providing custom LSI/VLSI, dedicated secure communications networks, fault tolerant computer architectures and computer network security.

SURVIVABLE and ENDURING SYSTEMS

With our primary focus on C^3I propagation and network analysis under benign and stressed environments, we're expanding modeling and assessment software tools to determine vulnerability, survivability, endurability and/or exploitation of C^3I systems worldwide.

Along with a network of project opportunities offering new growth comes a strong system of support and rewards. Success here is recognized within GTE and across the communications industry. Internal advancement can take the direction of management or development of technical knowledge. Our educational opportunities are unusually sophisticated and include an in-house MASTERS DEGREE program in Systems Engineering and Computer Science, and complete tuition reimbursement. Our benefits include life, health and dental insurance coverage for dependents. And because we provide year after year of challenge, CSD has kept top quality professionals and maintained low turnover in a fiercely competitive market for engineers. For access to the important opportunities in communications, we can be the last place you'll ever have to look.

We will also consider candidates who wish to enter Systems Engineering and have strong experience in our industry.

Send your resume, indicating areas of interest, to Greg N. Gostanian, GTE's Sylvania Systems Group, Communication Systems Division, 77 "A" Street, Needham, MA 02194. Or call us in Needham, MA at 1-800-225-3956.

An Equal Opportunity Employer M/F

If it's happening in communications, it's happening at CSD.

GTE Systems

mation that can be stored in a given area is multiplied by the number of colors that can be distinguished from one another at each site. An application of this principle is being studied by investigators at Cornell University, the International Business Machines Corporation, the General Motors Research Laboratories and Bell Laboratories. Instead of colored marks on paper the method exploits molecules that can effectively be given a color by laser irradiation. If the technique can be perfected, at least 1,000 colors might be distinguished at sites so small that 100 million of them could be packed into a square centimeter of surface area. Hence a storage density as high as 100 billion bits per square centimeter could be achieved.

A simpler method of optical recording for the bulk storage of information has already been demonstrated. A fine laser beam is directed onto a substrate and burns a small hole in the material. The presence or absence of a hole at a certain position encodes the information, and the size of the hole determines the storage density. The size is limited roughly by the wavelength of the laser radiation: the hole cannot be smaller than a square region one wavelength on a side. The technique now under study would combine the laser technology and a property of certain solids that was first described by Russian physicists in 1974 and has been investigated by workers at IBM since the late 1970's.

A solid can absorb radiation of a given frequency only if the molecules of the solid can resonate in response to the radiation and assume a higher energy state. In most circumstances the molecules return to the lower energy state as soon as the radiation is turned off; under some conditions, however, the molecules can undergo transformations that prevent them from falling back to their original state. Until recently it was thought that the molecular transformation must be a chemical one, in which the identity of the molecules is changed, but it has now been shown that the transformation can be a simple reorientation of the molecules. In both cases the effect is the same: the original, lower energy state is depopulated by the radiation, and the material ceases to resonate and absorb radiation at that frequency. (For visible radiation the frequency of the radiation corresponds to its color.) A persistent "hole" is created in the absorption spectrum of the material, and so the material becomes transparent across the narrow frequency band that corresponds to the resonant frequency of the original low-energy state.

Each hole in the absorption spectrum can correspond to one bit of information, and the spectrum available across each small region of a solid is wide enough to accommodate at least 1,000 spectral holes. All 1,000 bits can therefore be stored in an area one wavelength on a side. Investigation is now under way to enlarge the range of conditions and the family of materials that can be employed to make spectral holes.

Fire One

In man's reconstruction of his own prehistory an important development was the mastery of fire, which surely began with the capture and control of wildfire. In the 1930's evidence of hearths and charcoal dating to roughly 500,000 years ago was discovered at the *Homo erectus* site of Zhoukoudian, near Beijing. Hearths have since been found at Vértesszöllös in Hungary and at Tautavel in France, confirming that the knack of controlling fire was part of the behavioral repertory of *H. erectus.*

By the time of the European discoveries, however, *H. erectus* had been granted an antiquity substantially greater than 500,000 years. Remains of the species a full million years old had been found in the Far East and remains perhaps 1.5 million years old had been found in East Africa. If the species indeed had the cultural role of a Middle Paleolithic Prometheus (the Titan, it will be remembered, stole fire but did not make it), earlier evidence that *H. erectus* was able to control fire ought to exist.

The evidence may now have been found by a multidisciplinary team of investigators: J. W. K. Harris of the National Museums of Kenya, John Gowlett of the Research Laboratory for Archaeology and Art at the University of Oxford, Bernard Wood, an anatomist at the Middlesex Hospital Medical School in London, and Derek Walton of the department of physics at McMaster University. Working along the edge of the Rift Valley in the Laikipia Escarpment region of Kenya, they concentrated their efforts at a place called Chesowanja, on the east side of Lake Baringo, where artifacts and animal remains (including hominid remains) had already been found. The team soon discovered lumps of reddish mineralized matter. Walton, the geophysicist, determined that the lumps are composed of clay that has been baked to hardness at a temperature of 400 degrees Celsius. The distribution of the baked lumps indicated their direct association with stone artifacts, unmodified stone cobbles and fragmentary animal bones. At one site the baked material gave every appearance of being a manmade hearth.

Were the clay lumps formed by passing grass fires, a common event on the East African savanna? No; grass fires seldom raise the temperature of the ground above 65 degrees C. Might they be the result of exposure to more intense heat, such as smoldering tree stumps or strokes of lightning? Possibly; such chance events cannot be ruled out. Nevertheless, the association of the lumps of fired clay with hearth building, artifact manufacture and food consumption suggests that they are the result of human control of fire. How long ago was this? A chronology based on the radioactive decay of potassium into argon gives a date of about 1.4 million years ago. Who was tending the fire? Probably some early African example of *H. erectus.*

Coping with the Death Cap

Walter Litten, SCIENTIFIC AMERICAN's occasional correspondent on the "death cap" mushroom *Amanita phalloides* and other matters mycological, sends the following report:

"Even though poisoning by the death cap is not a bacterial disease and therefore seems unlikely to respond to antibiotics, evidence is accumulating that penicillin is a component of the most successful therapy for dealing with it. Two other components are oxygen-enriched air and silibinin, a substance from the milk thistle, *Silybum marianum.*

"A statistical analysis of 205 European cases of death-cap poisoning from 1971 through 1980 is presented in *Swiss Medical Weekly* by George L. Floersheim, O. Weber, P. Tschumi and M. Ulbrich. The three treatments were found to save more patients than were saved by any other therapy out of the 30 evaluated. The average number of therapies tried was eight, and some patients were subjected to as many as 20 therapies by physicians desperately 'trying everything.' Of the 205 patients, 45 died. This 22 percent mortality is no great advance over the 30 percent mortality recorded before the 1970's.

"Floersheim attributes the slight improvement not to better medical management but to greater awareness of the unreliability of traditional criteria for recognizing deadly species. He speculates that many of those who were poisoned ate sparingly of the mushrooms. A person who has swallowed much less of the toxins than is contained in one average-size specimen of *A. phalloides* has a good chance of recovery from the adventure in dining, given good supportive medical care.

"Floersheim states that the single most important factor distinguishing the survivors from the fatalities among the 205 patients is the quantity of *A. phalloides* ingested per unit of body weight. Analysis is complicated, however, by the considerable variation among mushrooms of the one species in their content of amatoxins, rings of eight amino acids that are considered principally responsible for the loss of liver function caused by the deadly amanitas. Recognizing these uncertainties, Floersheim and his colleagues nonetheless looked for a correlation between survival and the treatment administered.

"Thioctic acid, a relatively simple organic molecule, has been advanced as an antidote to amatoxin, and arrangements have been made for rushing it to physicians treating severe cases. In the 205 cases studied thioctic acid therapy turned out to be one of the measures taken more often with patients who did not survive. In the same category were treatment with the drug sulfamethoxazole, treatment of hemorrhaging induced by the toxin, plasma expanders to compensate for a reduced volume of circulating blood, hemodialysis to try to remove the toxins from the circulation and buffering to counter metabolic acidosis. For 20 other therapeutic measures the analysis showed neither positive nor negative correlation with outcome. On the positive side, penicillin and hyperbaric oxygen were found to be independently associated with a higher survival rate. The rate was enhanced when treatment with silibinin was combined with the injection of penicillin.

"One of several hypotheses to explain the apparent effect of penicillin in death-cap poisoning has been suggested by D. F. Schaefer and E. A. Jones in *The Lancet.* The proposed mechanism centers on gamma-aminobutyric acid (GABA), the chief inhibitor of nerve transmission. Bacteria in the gut synthesize GABA, and in the liver it is metabolized. A liver crippled by the attack of amatoxin does an inadequate job of keeping the GABA level in check. The result may be neural inhibition and the severe encephalopathy that is likely to be the final cause of death from *A. phalloides* and other amatoxin-laden toadstools. The role of penicillin may be to suppress the intestinal flora and so prevent the buildup of GABA until the liver mends.

"One of the later developments in severe *A. phalloides* intoxication is slowed blood clotting as the liver fails to manufacture the clotting factor thromboplastin. Clotting time is measured in terms of a percentage known as the Quick value. In Floersheim's survey 84 percent of the patients who reached a Quick value below 10 percent died; of the patients whose Quick value dropped no lower than 40 percent, none died."

Playing It by Ear

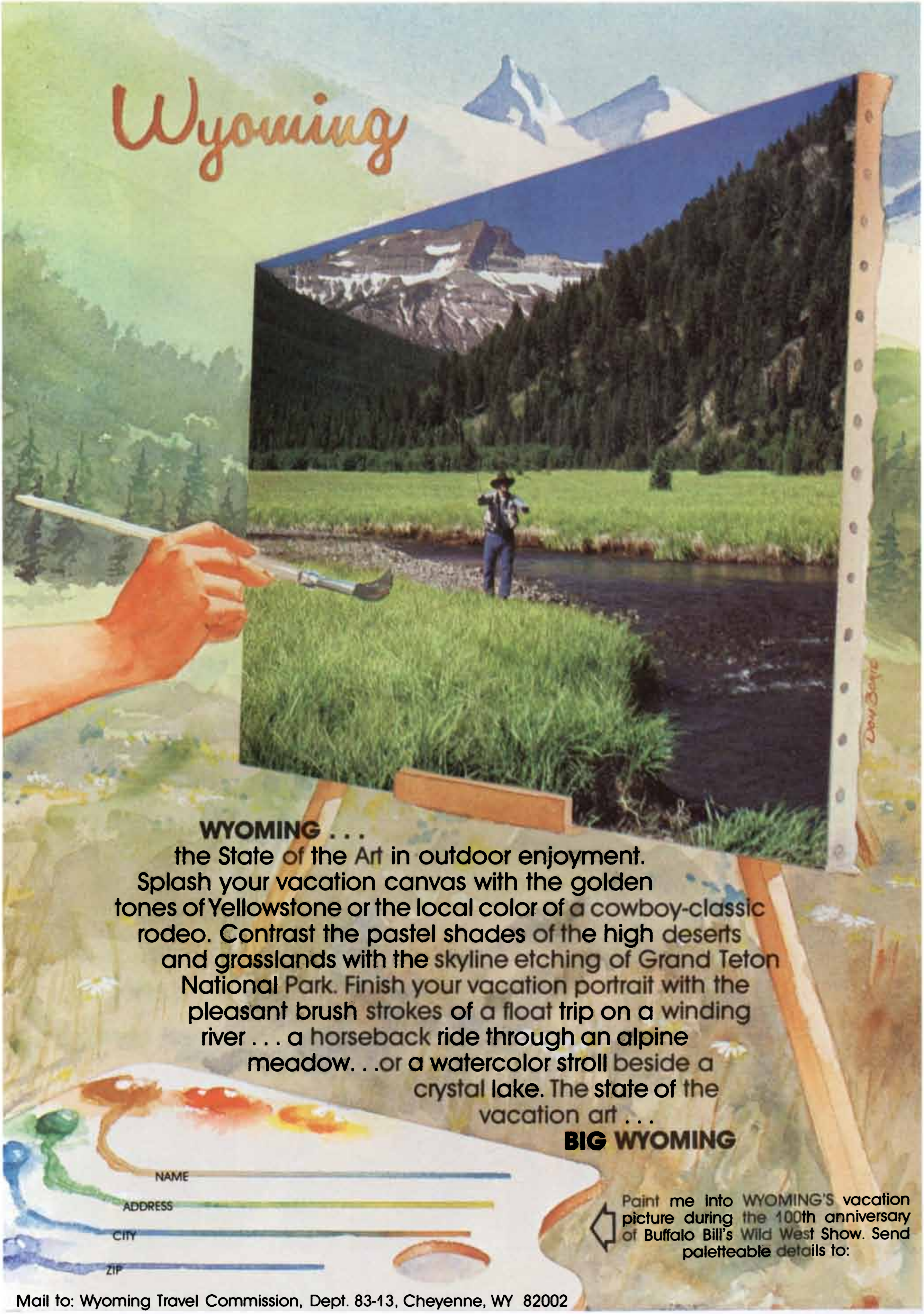
For a baseball outfielder to catch a fly ball is one thing; to explain how he does it is quite another. In 1968 Seville Chapman of the Cornell Aeronautical Laboratory considered the fly-ball question and advanced the hypothesis that the outfielder subconsciously solves a problem in trigonometry: he runs to a position on the field where the tangent of the ball's angle of elevation increases at a constant rate. Peter J. Brancazio of Brooklyn College of the City University of New York has reexamined the matter and proposes that the key is not trigonometry but the fielder's inner-ear sensors of motion. "We may actually be judging fly balls by ear," Brancazio says in summarizing a paper he presented at the joint annual meeting of the American Physical Society and the American Association of Physics Teachers.

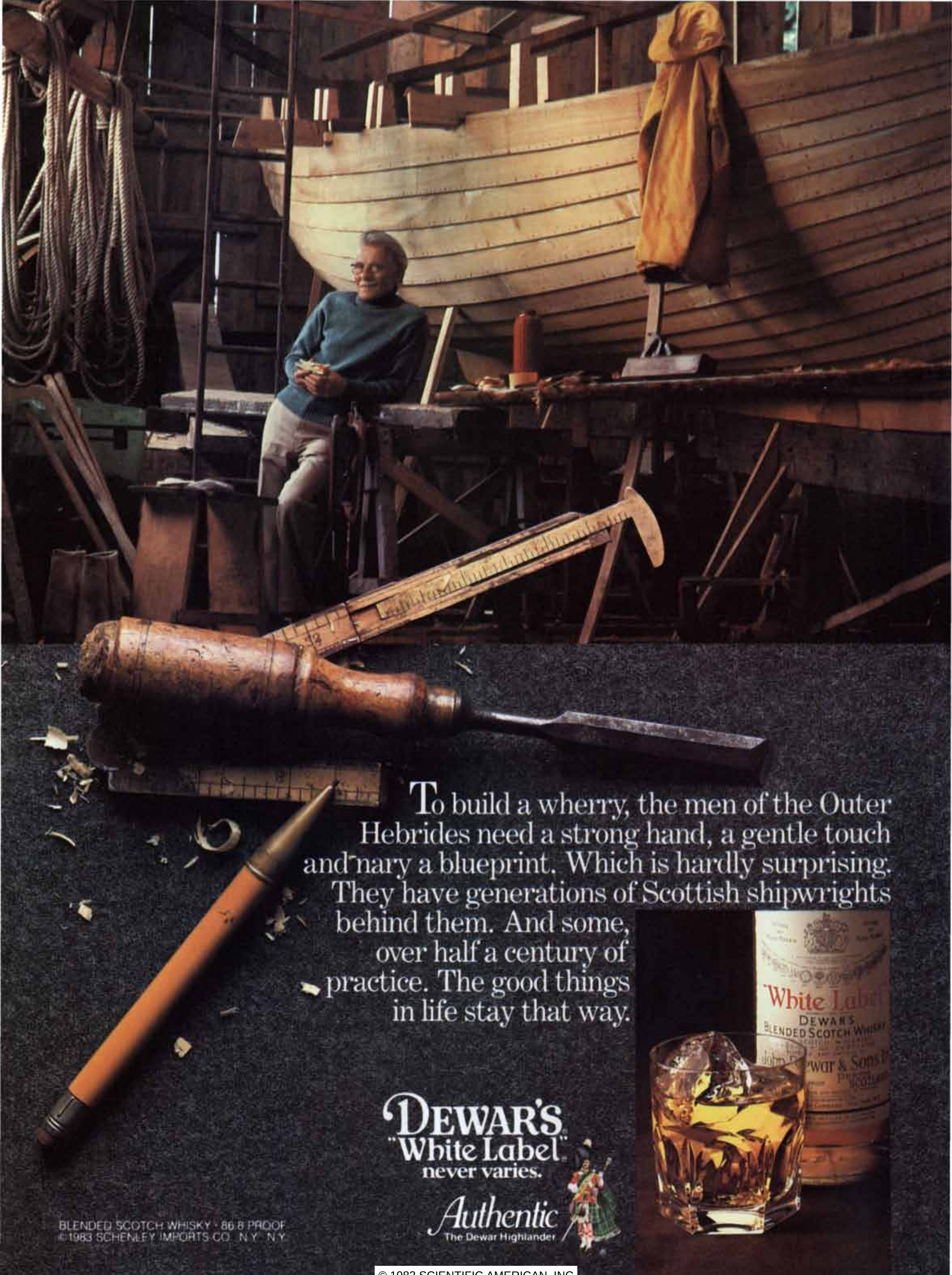
"One of the unique aspects of this skill," Brancazio writes, "is that there is no way to verbalize it or to teach someone how to do it. There do not seem to be any coaching techniques or helpful hints that could [improve] one's ability to judge fly balls accurately. The skill, it seems, must be completely self-taught on a nonverbal level."

In Chapman's view the outfielder teaches himself that if a ball is going over his head, he will see the tangent increasing at an increasing rate. (Brancazio calls this ball "Chapman's homer.") If the ball is going to fall in front of the fielder, he sees the tangent increasing at a decreasing rate. The fielder's art is one of learning where and how fast to run in order to keep the tangent increasing at a constant rate.

The difficulty with this hypothesis, Brancazio writes, is that it is based on the assumption that air resistance does not significantly affect the flight of the ball. "I have computed the effects of aerodynamic drag on the trajectory of a baseball and discovered that for the typical speeds and times of flight that occur under game conditions a batted baseball travels about 60 percent as far as it would in a vacuum. Moreover, air resistance distorts the shape of the trajectory of the ball so that it is noticeably different from a parabola. When aerodynamic forces are accounted for, the tangent of the angle of elevation does not increase at a constant rate as seen by a fielder who has judged the flight of the ball correctly."

Brancazio examined several other geometric and trigonometric features of the trajectory and found that none of them seems to show any characteristic variation that would tell the outfielder how to move to catch the ball. Considering the fly-ball problem as part of the larger question of how a person determines the position in space of any moving object and how he coordinates this information with the movement of his body, Brancazio notes that the person usually moves his head as well as his eyes as he follows the object. The motion of the head and that of the eyes must be closely coordinated. It turns out the coordination is guided primarily not by visual feedback but by signals triggered in the inner-ear sensors by the motions of the head. "This coordination of sensory input with body motion evidently follows a neural pathway that has been established through the familiar behavioral process of learning by trial and error."


Wyoming
WYOMING . . .
the State of the Art in outdoor enjoyment.
Splash your vacation canvas with the golden
tones of Yellowstone or the local color of a cowboy-classic
rodeo. Contrast the pastel shades of the high deserts
and grasslands with the skyline etching of Grand Teton
National Park. Finish your vacation portrait with the
pleasant brush strokes of a float trip on a winding
river . . . a horseback ride through an alpine
meadow. . .or a watercolor stroll beside a
crystal lake. The state of the
vacation art . . .
BIG WYOMING
NAME
ADDRESS
CITY
ZIP
Paint me into WYOMING'S vacation picture during the 100th anniversary of Buffalo Bill's Wild West Show. Send paletteable details to:
Mail to: Wyoming Travel Commission, Dept. 83-13, Cheyenne, WY 82002


To build a wherry, the men of the Outer Hebrides need a strong hand, a gentle touch and nary a blueprint. Which is hardly surprising. They have generations of Scottish shipwrights behind them. And some, over half a century of practice. The good things in life stay that way.
DEWAR'S
"White Label"
never varies.
Authentic
The Dewar Highlander
BLENDED SCOTCH WHISKY · 86.8 PROOF
©1983 SCHENLEY IMPORTS CO. N.Y., N.Y.